Pro OCR User's Guide

Contents

Chapter 1: Introducing Visioneer Pro OCR 100

Chapter 2: Learning Pro OCR Basics

Chapter 3: Getting Documents

Chapter 4: Locating Text and Graphics

Chapter 5: Setting Recognize Options and Proofing a

Pro OCR User's Guide

Chapter 1

Introducing Visioneer Pro OCR 100

This chapter introduces you to the Pro OCR application and to the concept of optical character recognition (OCR).

Why Pro OCR

Pro OCR is an Optical Character Recognition (OCR) application. An OCR application converts images of text, such as those obtained from scanning a document or receiving a fax through your fax-modem, into editable text. For example, when a scanner scans a page of text, it sees black and white areas on the page. The scanner converts what it sees into an image and stores the image on the computer. To transform a scanned text image into something a word processing or spreadsheet application can recognize as characters, you need an OCR (optical

Recognized Document

Chapter 6: Saving and Printing Documents

Chapter 7: Creating and Processing Deferred and Batch Jobs

Chapter 8: Tips for Getting the Best Results

Glossary

character recognition) application, such as Pro OCR.

Every day you may spend a lot of time retyping printed text or numbers from hard copy documents. By using Pro OCR and a scanner as an input device, you can eliminate much of this retyping.

Features and Highlights of Pro OCR

Many of the existing OCR products are typically capable of recognizing 200–300 plain, nonstylized typefaces. Using recognition technology, Pro OCR can recognize over 2,000 typefaces.

Most basic OCR applications inspect the scanned page image, attempt to recognize the dots on the page as characters, and transform the image into a plain text file. Pro OCR does all of these basic tasks, but it can also *get* the entire page into your word processor or spreadsheet as is—retaining the shape, form, type, and spacing, as well as the content, of the input page. Pro OCR provides:

- **The ability to read one or more pages of text including graphics.** Pro OCR reads pages directly from your scanner, or it reads TIFF, PCX, and DCX files. Pro OCR can automatically locate pictures and embed them in your document. You can also export pictures separately in a number of file formats.

- **Speed and accuracy of recognition.** With most documents, Pro OCR is faster than, and as accurate as a good typist.

- **Numeric regions.** You can specify that a given region on a page can contain only numbers. Numeric regions help Pro OCR make sure that numbers are always recognized as numbers and never mistakenly identified as

letters.

- **Recognition and retention of fonts, characters, styles, and page formatting.** Pro OCR recognizes and retains the differences between serif and sans-serif fonts, styles such as bold, underline, and subscript, and formatting such as columns, tables, and indents.

- **Deferred and batch processing.** You can perform procedures that need your attention or interaction (for example, locating), and then do the time consuming steps that don't need interaction (for example, recognizing) at another time.

- **Internet readiness.** supports HTML export format. You can convert an image file directly to an HTML page and upload it to the Web site.

- **Proofing options.** Pro OCR has a number of proofing options. You can also send recognized text directly to your word processor.

- **Save features.** With Pro OCR you can save recognized text in a wide variety of word processor and spreadsheet file formats. Pro OCR works with imperfect input pages that may have skewed lines of text, touching or broken characters, and fuzzy characters.

© Copyright 1998 Visioneer, Inc. Reach us at www.visioneer.com.

Pro OCR User's Guide

Chapter 1

Introducing Visioneer Pro OCR 100

This chapter introduces you to the Pro OCR application and to the concept of optical character recognition (OCR).

Why Pro OCR

Pro OCR is an Optical Character Recognition (OCR) application. An OCR application converts images of text, such as those obtained from scanning a document or receiving a fax through your fax-modem, into editable text. For example, when a scanner scans a page of text, it sees black and white areas on the page. The scanner converts what it sees into an image and stores the image on the computer. To transform a scanned text image into something a word processing or spreadsheet application can recognize as characters, you need an OCR (optical character recognition) application, such as Pro OCR.

Every day you may spend a lot of time retyping printed text or numbers from hard copy documents. By using Pro OCR and a scanner as an input device, you can eliminate much of this retyping.

Features and Highlights of Pro OCR

Many of the existing OCR products are typically capable of recognizing 200–300 plain, nonstylized typefaces. Using recognition technology, Pro OCR can recognize over 2,000 typefaces.

Most basic OCR applications inspect the scanned page image, attempt to recognize the dots on the page as characters, and transform the image into a plain text file. Pro OCR does all of these basic tasks, but it can also *get* the entire page into your word processor or spreadsheet as is—retaining the shape, form, type, and spacing, as well as the content, of the input page. Pro OCR provides:

- **The ability to read one or more pages of text including graphics.** Pro OCR reads pages directly from your scanner, or it reads TIFF, PCX, and DCX files. Pro OCR can automatically locate pictures and embed them in your document. You can also export pictures separately in a number of file formats.

- **Speed and accuracy of recognition.** With most documents, Pro OCR is faster than, and as accurate as a good typist.

- **Numeric regions.** You can specify that a given region on a page can contain only numbers. Numeric regions help Pro OCR make sure that numbers are always recognized as numbers and never mistakenly identified as letters.

- **Recognition and retention of fonts, characters, styles, and page formatting.** Pro OCR recognizes and retains the differences between serif and sans-serif fonts, styles such as bold, underline, and subscript, and formatting such as columns, tables, and indents.

- **Deferred and batch processing.** You can perform procedures that need your attention or interaction (for example, locating), and then do the time consuming steps that don’t need interaction (for example, recognizing) at another time.

- **Internet readiness.** supports HTML export format. You can convert an image file directly to an HTML page and upload it to the Web site.

- **Proofing options.** Pro OCR has a number of proofing options. You can also send recognized text directly to your word processor.

- **Save features.** With Pro OCR you can save recognized text in a wide variety of word processor and spreadsheet file formats. Pro OCR works with imperfect input pages that may have skewed lines of text, touching or broken characters, and fuzzy characters.

© Copyright 1998 Visioneer, Inc. Reach us at www.visioneer.com.

Copyright Information

Pro OCR User's Guide for Windows. Copyright ©1998 Visioneer, Inc. All rights reserved.

Reproduction, adaptation, or translation without prior written permission is prohibited, except as allowed under the copyright laws.

AnyPort, AutoFix, AutoLaunch, FormTyper, MicroChrome, PaperEnable, PaperLaunch, PaperPort, PaperPort Deluxe, PaperPort ix, PaperPort Links, PaperPort mx, PaperPort PowerBar, PaperPort 3000, PaperPort 6000, PaperPort vx, PaperPortation, PaperPort Strobe, Pro OCR, ScanDirect, SimpleSearch, SharpPage, and Visioneer are trademarks of Visioneer, Inc. PaperPort, Paper-driven, and the Visioneer logo are registered trademarks of Visioneer, Inc.

Microsoft is a U.S. registered trademark of Microsoft Corporation. Windows is a trademark of Microsoft Corporation. TextBridge is a registered trademark of Xerox Corporation. ZyINDEX is a registered trademark of ZyLAB International, Inc. ZyINDEX toolkit portions, Copyright © 1990–1996, ZyLAB International, Inc. All Rights Reserved. All other products mentioned herein may be trademarks of their respective companies.

Information is subject to change without notice and does not represent a commitment on the part of Visioneer, Inc. The software described is furnished under a licensing agreement. The software may be used or copied only in accordance with the terms of such an agreement. It is against the law to copy the software on any medium except as specifically allowed in the licensing agreement. No part of this document may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or information storage and retrieval systems, or translated to another language, for any purpose other than the licensee's personal use and as specifically allowed in the licensing agreement, without the express written permission of Visioneer, Inc.

Part Number: 05-0340-000

Restricted Rights Legend

Use, duplication, or disclosure is subject to restrictions as set forth in contract subdivision (c)(1)(ii) of the Rights in Technical Data and Computer Software Clause 52.227-FAR14. Material scanned by this product may be protected by governmental laws and other regulations, such as copyright laws. The customer is solely responsible for complying with all such laws and regulations.

Visioneer's Limited Product Warranty

If you find physical defects in the materials or the workmanship used in making the product described in this document, Visioneer will repair, or at its option, replace, the product at no charge to you, provided you return it (postage prepaid, with proof of your purchase from the original reseller) during the 12-month period after the date of your original purchase of the product.

THIS IS VISIONEER'S ONLY WARRANTY AND YOUR EXCLUSIVE REMEDY CONCERNING THE PRODUCT, ALL OTHER REPRESENTATIONS, WARRANTIES OR CONDITIONS, EXPRESS OR IMPLIED, WRITTEN OR ORAL, INCLUDING ANY WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE EXPRESSLY EXCLUDED. AS A RESULT, EXCEPT AS SET OUT ABOVE, THE PRODUCT IS SOLD "AS IS" AND YOU ARE ASSUMING THE ENTIRE RISK AS TO THE PRODUCT'S SUITABILITY TO YOUR NEEDS, ITS QUALITY AND ITS PERFORMANCE,

IN NO EVENT WILL VISIONEER BE LIABLE FOR DIRECT, INDIRECT, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES RESULTING FROM ANY DEFECT IN THE PRODUCT OR FROM ITS USE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

All exclusions and limitations in this warranty are made only to the extent permitted by applicable law and shall be of no effect to the extent in conflict with the express requirements of applicable law.

FCC Radio Frequency Interference Statement

This equipment has been tested and found to comply with the limits for the class B digital device, pursuant to part 15 of the FCC Rules. These limits are designed to provide reasonable protection against interference in a residential installation. This equipment generates, uses and can radiate radio frequency energy and if not installed, and used in accordance with the instructions, may cause harmful interference to radio communications. However, there is no guarantee that interference will not occur in a particular installation. If this equipment does cause harmful interference to radio or television reception, which can be determined by turning the equirpment off and on, the user is encouraged to try and correct the interference by one or more of the following measures:

- Reorient or relocate the recemng antenna.

- Increase the separation between the equipment and receiver.
- Connect the equipment into an outlet on a circuit different from that to which the receiver is connected.
- Consult the dealer or an experienced radio/TV technician for help.

This equipment has been certified to comply with the limits for a class B computing device, pursuant to FCC Rules. In order to maintain compliance with FCC regulations, shielded cables must be used with this equipment. Operation with non-approved equipment or unshielded cables is likely to result in interference to radio and TV reception. The user is cautioned that changes and modifications made to the equipment without the approval of manufacturer could void the user's authority to operate this equipment.

This device complies with part 15 of the FCC Rules. Operation is subject to the following two conditions: (1) This device may not cause harmful interference, and (2) this device must accept any interference received, including interference that may cause undesired operation.

Back to Pro OCR User’s Guide.

Contents

Chapter 1: Introducing Visioneer Pro OCR 100

Chapter 2: Learning Pro OCR Basics

Chapter 3: Getting Documents

Chapter 4: Locating Text and Graphics

Chapter 5: Setting Recognize Options and Proofing a Recognized Document

Chapter 6: Saving and Printing Documents

Chapter 7: Creating and Processing Deferred and Batch Jobs

Chapter 8: Tips for Getting the Best Results

Glossary

Contents

Chapter 1: Introducing Visioneer Pro OCR 100

Why Pro OCR

Features and Highlights of Pro OCR

Glossary

Glossary

A4 Letter page size

accelerator key

ADF

alphanumeric word

ASCII

As Single Column locating method

Auto OCR

Auto brightness

automatic document feeder (ADF)

automatic processing

background noise

backup

backwards compatible

bit image

bitmap

bitmapped character

bold text

brightness

broken character

built-in dictionary

CCITT

character

character format

character identification error

character image

character recognition

character style

clipboard

column information

compression

confidence

consistent document

copyrighted document

deferred job

deferred processing

degraded image

dialog box

desktop

document area

dots per inch (dpi)

dpi

draft quality text

driver

exporting

export format

file extension

file formats

file type

fine resolution

flatbed scanner

font

font family

font mapping

format retention

Gallery

Get Page

grayscale image

hard page breaks

heavy character

I-beam pointer

icon

illegible character

illegible character symbol

image view

input file formats

insertion point

italic text

justification

kerning

landscape orientation

layout

layout analysis error

Legal page size

Lenient suspect threshold

letter quality text

line break

Locate

locate region

locating

locating method

menu

menu bar

multi-column text

monospaced font

monospaced font mapping

newspaper style columns

Normal locating method

Normal suspect threshold

numeric region

OCR

On-Screen Verifier™

Optical Character Recognition (OCR)

order of text regions

orientation

output file formats

page controls

page format

page image

page number box

page orientation

page size

page source

PCX

picture element

picture region

pixel

pixel-for-pixel

plain text

portrait orientation

printer font

Pro OCR Deferred format

Pro OCR format

Pro OCR process

Pro OCR window

Proof

proportionally spaced font

recognition accuracy

Recognize

recognized text

recognizing

region style

resolution

Rich Text Format (RTF)

RTF

sans serif

sans serif font mapping

scanner

scanner driver

scanning

screen font

scroll bars

serif

serif font mapping

settings file

sheetfed scanner

side-by-side columns

single-bit image

single-step processing

skewed text

spell checking

standard resolution

status bar

status display area

Stringent suspect threshold

stroke weight

Style ribbon

stylized font

subscript text

superscript text

supplementary dictionaries

suspect character

suspect threshold

Tag Image File Format

template

template matching

Template locating method

text quality

text region

text style

text view

throughput

TIFF

touching characters

typeface

type quality

type size

type style

underline text

User Defined page size

user dictionary

view selector

window

Windows

word wrap

zoom controls

Glossary

A4 Letter page size	An A4 size page measures 8.33" x 11.66".
accelerator key	In Windows applications, a keyboard shortcut to a menu command.
ADF	See **automatic document feeder (ADF)**.
alphanumeric word	A word made up of the alphabetic and numeric characters (A–Z, a–z, 0–9) in a character set. Excludes punctuation and other symbol characters.
ASCII	Acronym for American Standard Code for Information Interchange (pronounced “ASK-ee”). A standard that assigns a unique binary number to each text and control character. ASCII code is used for representing text inside a computer and for transmitting text between computers or between a computer and a peripheral device.
As Single Column locating method	One of Pro OCR’s three locating methods. Use it when you want Pro OCR to read a page as a single column, from left margin to right margin, ignoring any column or paragraph spacing. Most commonly used for pages in which there is no clear column or paragraph structure.
Auto OCR	Clicking this button starts automatic processing, which uses Get Page, Locate, and Recognize according to the current gallery settings.
Auto brightness	A feature of some scanners, by which brightness is adjusted automatically while the page is scanned.
automatic document feeder (ADF)	Built-in or optional equipment for a scanner that lets you automatically scan stacks of pages instead of having to place them one at a time on the flatbed. Sometimes it’s difficult to control the proper alignment of pages using an automatic document feeder. Compare with **flatbed scanner** and **sheetfed scanner**.

automatic processing	A method for using Pro OCR with minimal intervention. Automatic processing involves setting appropriate Gallery settings, before using Auto Start to read in one or more image files or scan in one or more pages. Once page images have been acquired, automatic processing Locates and Recognizes each page image in succession. Automatic processing is best suited to documents that require the same Gallery settings (Page Size, Brightness, Locate method, etc.). Compare with **single-step processing**.
background noise	Non-character or non-graphic information in a page image that adversely affects optical recognition. Background noise includes the shading that results from scanning colored paper stock, extraneous marks, dirt or ink bleed. Problems with background noise can be reduced by using the brightness setting in Pro OCR to compensate for the type of noise on the page.
backup	(n.) A copy of a disk or of a file on a disk. It's a good idea to make backups of all your important disks and to use the copies for everyday work, keeping the originals in a safe place.
backwards compatible	The ability of an application to open files created with earlier versions of that application.
bit image	A collection of bits in memory that represents a two-dimensional surface. For example, the screen is a visible bit image.
bitmap	**1.** A set of bits that represents the graphic image of an original document in memory. **2.** A set of bits that represents the positions and states of a corresponding set of items, such as pixels. Used by the computer to construct graphic images and fonts. See also **bit image**.

bitmapped character	A character image made up of a pattern of dots that exists in a computer file or in memory as a bitmap. Bitmapped characters cannot be interpreted by a computer. In order for a computer to use bitmapped characters in a word processor or spreadsheet, the characters must first be interpreted by an OCR application and translated into ASCII text.
bold text	Text with the **bold attribute looks like this**. See also **text style**.
brightness	The relative amount of light or darkness reflected from an image. A scanner's brightness control is used in Pro OCR to adjust for pages that are either too light or too dark.
broken character	A character with one or more missing pieces, such as a missing serif, stem, or cross bar. For example, a broken lower case 'e' might not have a fully closed loop, which could cause it to be misrecognized. Problems with broken characters can be reduced by using the brightness setting in Pro OCR to darken the image when scanning. Compare with **heavy character** and **touching characters**.
built-in dictionary	The dictionary that Pro OCR automatically loads and uses whenever Recognize is done. The built-in dictionary is used to enhance Pro OCR's recognition accuracy and also to find misspelled words in the document. Compare with **supplementary dictionaries** and **user dictionary**.
CCITT	Abbreviation for *Consultative Committee on International Telegraphy and Telephony*; an international committee that sets standards and makes recommendations for international communication. One of the standards set by CCITT is for the compression of image files. Pro OCR employs CCITT-standard compression methods. See also **compression** and **TIFF**.
character	Any symbol that has a widely understood meaning and thus can convey information, including alphabetic, numeric, symbolic, and punctuation elements.

character format	Font and style information applied to characters. Character format information includes the font name and type size, as attributes such as underline, bold, italic, or some combination of these properties. Compare with **page format**.
character identification error	An incorrectly recognized bitmapped character. There are two kinds of character identification errors—substitutions and rejects. A character substitution occurs when a character is incorrectly recognized as another. A reject character results from the inability of the OCR application to interpret a character image with sufficient confidence. In such cases, recognition is not attempted and the character is flagged as illegible. Compare with **layout analysis error**.
character image	An arrangement of bits that defines a character in a font.
character recognition	The OCR process in which bitmapped character images are interpreted and translated into ASCII computer codes.
character style	See **type style**.
clipboard	In Windows applications, temporary storage for text that is cut or copied from a document. Text saved in the clipboard may be pasted back into the same or another document.
column information	Part of Pro OCR's page format information. Column information includes the location of the column on the page, the width of the column, and its left and right margins.
compression	Electronic method for reducing the size of a file without losing any information in the file. Compressed TIFF files take up significantly less disk space than uncompressed files. See also **TIFF** and **CCITT**.
confidence	In Pro OCR, a measure of the certainty of an unknown character's identity. Above a certain confidence level, a character is automatically recognized. At lower confidence levels, a character may either be recognized, but flagged as a suspect character, or not recognized and flagged as an illegible character.

consistent document	A set of pages or image files where the same Gallery settings apply to each page in the document. Pro OCR's Auto Start feature can be used to best effect when a document is consistent.
copyrighted document	Most published or printed materials and documents are copyrighted. It is illegal to use a computer and Pro OCR to copy, store, or reproduce, on paper or electronically, any copyrighted documents without the permission of the copyright holder.
deferred job	A file that contains one or more partially processed pages for Pro OCR to finish processing later on. See also **Pro OCR Deferred format** and **deferred processing**.
deferred processing	Provides the ability to individually specify Get Page, Locate, and recognize settings for particular pages when necessary, while still being able to automatically process a job at a later time.
degraded image	An image that contains broken characters, touching characters and/or background noise. See **broken character**, **touching characters** and **background noise**.
dialog box	In Windows applications, the standard pop-up box that is displayed to communicate with the user when a command requires some further action. Some dialog boxes are informational.
desktop	Your working environment on the computer—the menu bar and the background area on the screen. You can have a number of documents or windows on the desktop open at the same time.
document area	The main part of the application window in Pro OCR. The document area shows one page of the current document at a time using the selected View Size setting.

dots per inch (dpi)	A measure of the visual resolution of a display or output device. Monitor screens typically have resolutions in the range of 70 to 75 dpi. Most common laser printers have a resolution of 300 dpi. The lower the resolution of a page in dots per inch, the lower the visual quality of characters on that page. Pro OCR can quickly and accurately recognize characters scanned in at resolutions down to 200 dpi.
dpi	See **dots per inch (dpi)**.
draft quality text	On 9-pin dot matrix printers, the low resolution printing option. Draft quality text is monospaced and made up of visible dots that do not touch. In Pro OCR, click the Draft Quality button in the Recognize section of the Gallery, to improve recognition on draft quality dot matrix text. Compare with **letter quality text**.
driver	See **scanner driver**.
exporting	Saving a document in an external format, such as a word processor, spreadsheet, text or standard image file. An exported document is created for use outside of Pro OCR.
export format	Pro OCR can save and export documents in a variety of specific word processor and spreadsheet formats. The specific export format is specified in the Save As dialog box.
file extension	In the MS-DOS operating system, file names conventionally consist of a base and a file extension, for example `SAMPLE.TXT`. In this example, "`SAMPLE`" is the base, and the file extension is "`.TXT`". File extensions are used to identify the type of file. In this example, the file extension indicates that this is a text (ASCII) file.
file formats	See **input file formats** and **output file formats**.
file type	Different applications create different file types. Some file types are application-specific. Other file types are generic. The file type indicates what kind of information is contained in the file and what format the information is in. Most applications can only open files of certain file types.

fine resolution	A term associated with FAX modems, referring to the highest resolution of the image files typically produced by these devices. Fine resolution is approximately 200 x 200 dpi, which is adequate for reliable recognition.
flatbed scanner	Scanner with a glass plate on which pages are placed face down. Although such scanners can only read one page at a time, they can support a variety of paper sizes and it's easier to control the proper alignment of a page. Compare with **automatic document feeder (ADF)** and **sheetfed scanner**.
font	All characters (letters, numbers, and symbols) in one size and style of a font family. 12 point Helvetica Bold Italic is a font. "Font" is sometimes incorrectly used instead of "font family" or "typeface." See also **font family** and **typeface**.
font family	The complete set of variations of a particular typeface. For example, Helvetica is a font family. It contains a variety of typefaces including, for example, Helvetica, Helvetica Bold, Helvetica Italic, Helvetica Bold Italic. See also **font** and **typeface**.
font mapping	Set in the Display Options dialog box. Tells Pro OCR which fonts to use to display recognized text. Also specifies which fonts to use in documents that are exported to Windows-based word processors.
format retention	The ability to retain the layout of a page, including margins, paragraph and column widths, and tabs and indents. Pro OCR preserves as much page format information as export formats support.
Gallery	The Pro OCR toolbar. All settings for the Get Page, Locate, and Recognize stages of the Pro OCR process are set in the Gallery. Common Pro OCR processes—Auto Start, and single-step Get Page, Locate, and Recognize—can be initiated from the Gallery.
Get Page	Single-step Gallery function. It is also the first stage of the Pro OCR process. Scans one page from a scanner or reads one file, using the current Get Page settings.

grayscale image	An image format where individual pixels can be expressed with more than a single bit, allowing the image to contain true shades of gray. Pro OCR will not open grayscale images. Compare with **single-bit image**.
hard page breaks	Special formatting that you put in manually in a text or word processor document. Most word processors and text editors automatically create soft page breaks unless you explicitly specify hard page breaks. In Pro OCR, you can force the output application to preserve the page breaks of the input document by clicking the “Insert Hard Page Breaks” checkbox when you are in the Save As Options dialog box.
heavy character	In Pro OCR, a character that is printed too dark or thick, so that the representation obscures detail and reduces confidence in the identity of that character.
I-beam pointer	A mouse pointer shape that resembles an upper-case “I”. When the pointer has this shape, you can select text. See also **insertion point**.
icon	An image that graphically represents an object, a concept, or a message. Screen icons can represent disks, documents, application programs, or other things you can select and open. In an application such as Pro OCR, icons are also used to represent various settings in the gallery, Style ribbon, and Status bar.
illegible character	A character that Pro OCR cannot recognize with adequate certainty. Illegible characters in a document are highlighted and displayed with the specified illegible character symbol in the text view. See also **suspect character**.
illegible character symbol	The symbol Pro OCR uses to display illegible characters in the text view. Set in the Display Preferences dialog box. See also **illegible character**.
image view	The view that displays the bitmapped image of a page. Used to locate regions of text or graphics, and for viewing the original scanned image of a page during proofing and editing.

input file formats	Pro OCR can read documents saved by other applications in TIFF, PCX and DCX formats, as well as those documents saved in its own proprietary TIFF format. See also **PCX** and **TIFF**.
insertion point	The place in a text file where text is inserted or deleted. Indicated by a blinking vertical bar.
italic text	Text with the *italic attribute looks like this*. See also **text style**.
justification	Alignment of text to the left, right, or both margins of a column or page. Text may be left-justified, right-justified, center-justified, or fully justified (both left- and right-justified). Pro OCR preserves justification.
kerning	A measure of the spacing between characters. In tightly kerned text, the letters are very close together, which can cause letters to touch when the page is scanned. See also **touching characters**.
landscape orientation	When you hold a page of text to read it, it is in landscape orientation when the page is wider than it is tall. Compare with **portrait orientation**.
layout	The relative position of elements on a page, such as margins, columns, graphics, titles and sections.
layout analysis error	The result of an OCR product's inability to correctly organize recognized text into words, lines and paragraphs on the page. There are two kinds of common layout analysis errors—incorrectly interpreting the flow of text on a page and incorrectly grouping or separating side-by-side paragraphs. Layout analysis errors can be more troublesome than character identification errors, particularly with documents having complex layouts. Compare with **character identification error**.
Legal page size	See **page size**.

Lenient suspect threshold	Tells Pro OCR to only highlight suspect characters it is very uncertain of. Very few characters are marked as suspect, compared to when the suspect threshold is set to normal or stringent. Use it when you're dealing with documents containing fonts that you know from experience have been recognized accurately or when you're less concerned with double-checking. Set in the Display Options dialog box. Compare with **Normal suspect threshold** and **Stringent suspect threshold**.
letter quality text	Text made up of characters that are fully formed with dots that are touching. Compare with **draft quality text**.
line break	The point at the edge of a line of text where the text flows onto the next line.
Locate	Single-step Gallery function. It is also the second stage of the Pro OCR process. Specifies which text will be recognized on a page by creating or applying locate regions on the page according to the current Locate and Pictures settings. The current Locate setting may be either Normal, As Single Column, or Template. The current Pictures setting may be Locate Text and Pictures or Locate Text Only.
locate region	Defines an area on the page image in the image view and the text view. The text and picture kinds of locate regions may be defined automatically or manually. All three types of locate regions may be manually defined using the locate region drawing feature, or may be recalled using the Template locating method. See also **text region**, **numeric region**, **picture region**, and **Template locating method**.
locating	The process in Pro OCR for specifying which locate regions will be recognized on a page by creating or applying locate regions on the page.
locating method	Tells Pro OCR how to locate regions for processing on a page. The three locating methods are Normal, As Single Column, and Template. See also **Normal locating method**, **As Single Column locating method**, and **Template locating method**.

menu	A list of choices from which the user can choose. Menus appear when you point to and click a menu title in the menu bar, or a pop-up menu title in a window or dialog box.
menu bar	The horizontal strip at the top of a window that contains menu titles.
multi-column text	Text that is formatted into more than one column on a single page. Examples include phone books and newspapers.
monospaced font	Also known as a fixed pitch font. A typeface, such as Courier, in which each character takes up the same amount of horizontal space. The output from most typewriters is monospaced. Compare with **proportionally spaced font**.
monospaced font mapping	The font chosen for displaying monospaced text characters in text views. Set in the Display Options dialog box. Compare with **sans serif font mapping** and **serif font mapping**.
newspaper style columns	Also known as “snaked” or winding columns. A column format where the text flows down the vertical length of the column before moving to the top of the next column. As the name suggests, this type of column is commonly found in newspaper and magazine articles. This glossary is formatted in newspaper style columns. The flow of text in newspaper style columns is best suited for the Normal locate setting in Pro OCR.
Normal locating method	One of Pro OCR’s three locating methods. Use it for most kinds of input, including many tables and forms. Creates text regions based on column or paragraph spacing. Compare with **As Single Column locating method** and **Template locating method**.
Normal suspect threshold	Tells Pro OCR to highlight suspect characters that it is somewhat uncertain of. More characters are marked as suspect than when a lenient suspect threshold is used. Use it with clean, clear, typeset documents when most of the words in the document are probably in the dictionaries. Set in the Display Options dialog box. Compare with **Lenient suspect threshold** and **Stringent suspect threshold**.

numeric region	Defines a numeric area on the page image in Image View and Text View. Numeric regions may be defined using Pro OCR's manual region drawing feature, or may be recalled using the Template locating method. Compare with **text region** and **picture region**. See also **Template locating method**.
OCR	See **Optical Character Recognition (OCR)**.
On-Screen Verifier™	Pops up in the document area to display a section of the page image corresponding to the current text selection in the text view. The on-screen verifier is displayed automatically when proofing, and can also be shown or hidden by choosing the Show/Hide On-Screen Verifier command from the Edit menu.
Optical Character Recognition (OCR)	The process by which a computer converts scanned text images into editable text characters.
order of text regions	Shown by an arrow from the center of a text region to the top center of the next text region, in Image View after Locating has been done. Text is output to application files in the order in which text regions are specified.
orientation	Determines the angle or rotation of the page. Pro OCR allows you to choose between portrait or landscape orientation. See also **portrait orientation** and **landscape orientation**.
output file formats	Pro OCR can save documents in a variety of formats, including ASCII, a multitude of export formats, the Pro OCR format, and Pro OCR Deferred format. See also **export format**, **Pro OCR format**, and **Pro OCR Deferred format**.
page controls	Contains the previous and next page arrows and the page number box. Click the previous page arrow or the next page arrow to move from page to page in a document. See also **page number box**.
page format	The layout of the page, including its margins, paragraph and column widths, and tabs and indents. Pro OCR preserves nearly all page format information. What page format information is preserved in saved application files depends on the application format.

page image	The bitmapped image of a scanned page, displayed in the image view in Pro OCR.
page number box	Shows which page is being viewed and how many pages are in the document. Double-click it to go to a specific page. See also **page controls**.
page orientation	See **orientation**.
page size	The width and height to use when getting a page from a scanner within Pro OCR. There are three pre-defined page sizes: US Letter, US Legal, and A4 Letter. There is also an option for user-defined page sizes.
page source	Pro OCR can get pages from a file or the selected scanner. You can draw pages from either source at any time.
PCX	A common graphic file format on MS-DOS computers. Some scanners produce PCX files. Pro OCR can read single PCX files produced by many scanners, fax cards, and graphics applications. A variation of the PCX format is DCX—a multi-page PCX file. Pro OCR can also read DCX files.
picture element	See **pixel**.
picture region	Defines a picture area on the page image. Picture regions may be defined manually or by using the Locate button with “Locate Text and Pictures” selected.
pixel	A single unit (or dot) of screen, printer or image resolution. The number of pixels (or dots) per inch determines the resolution of an image. Most scanners and laser printers offer resolutions of at least 300 pixels (or dots) per inch.
pixel-for-pixel	A large magnified image view (approximately 400%) of the page. Lets you inspect the quality of the image. Each screen pixel corresponds to one image pixel.
plain text	Text with no special attributes or styling, such as **bold**, *italic*, or underline.

portrait orientation	When you hold a page of text to read it, it is in portrait orientation when the page is taller than it is wide. Compare with **landscape orientation**.
printer font	The representation of a font or typeface used for printing by a printer. See also **font**, **font family**, and **typeface**. Compare with **screen font**.
Pro OCR Deferred format	One of Pro OCR's output file formats. Saves a document with the current state of Get Page, Locate, and Recognize for every page. When the document is processed using Process Deferred Job, the saved information is retained and only those processes and pages that have not already been specified are completed using the current Gallery settings.
Pro OCR format	Pro OCR's native/internal file format. The Pro OCR format is a proprietary variation of the Group 4 TIFF format. Documents at various stages of processing may be saved in this format and opened later for additional processing.
Pro OCR process	The five stage process that translates printed text or image files into an output form suitable for use in other applications. The five steps of the Pro OCR process are: Get Page, Locate, Recognize, Proof/Edit and Save/Export.
Pro OCR window	The main window for interacting with Pro OCR. Contains the title bar, menu bar, gallery, scroll bars, Status bar, and document area.
Proof	The fourth stage in the Pro OCR process, where any suspect and illegible characters or misspelled words can be examined and corrected, if necessary. This command moves the insertion point to the next piece of text in the text view, according to the Proofing Options. The Proofing Options configure Proof to view suspect or illegible characters, misspelled words, punctuation, numbers, alphanumeric words, or entire lines at a time. Use the Tab key as a keyboard shortcut.

proportionally spaced font	Also known as a variable pitch font. Typeface in which each character takes up an amount of horizontal space consistent with its relative physical width, i.e. an “i” needs less space than a “w.” Times Roman and Helvetica are two common proportionally spaced typefaces. Compare with **monospaced font**.
recognition accuracy	A measure of the degree to which OCR output conforms to the individual characters in the input document. Recognition accuracy is a percentage expression of the number of correct character identifications in relation to the total number of characters in the page or document. This measure is often used as the primary criterion in evaluating OCR performance, even though it does not account for layout analysis errors. Compare with **throughput**.
Recognize	Single-step Gallery function. It is also the third stage of the Pro OCR process. The process in Pro OCR in which bitmapped text images are converted into editable text. Recognizes text defined by the text regions on the current page according to the current Recognize setting.
recognized text	The initial result of OCR processing. Once an image has been recognized, the resultant text can be proofed/edited and exported to other applications.
recognizing	The process in Pro OCR in which character images are converted into digital computer character codes (ASCII equivalents).
region style	The type of a locate region, either text, numeric or picture. See also **locate region**, **text region**, **numeric region**, and **picture region**.
resolution	Density of pixels in an output device such as a screen display or printer, or in an input device such as a scanner. Usually specified in dots per inch. See also **dots per inch (dpi)**.
Rich Text Format (RTF)	An output file format for word processors that preserves most page format and font information. One of Pro OCR’s export file formats. Many Windows-based word processors can read files in RTF, although they have varying levels of support for RTF.

RTF	See **Rich Text Format (RTF)**.
sans serif	Designation for font families in which the characters do not have serifs, which are the small strokes at the ends of characters. Common sans serif font families include Helvetica, Avant Garde, and Univers. Compare with serif.
sans serif font mapping	The font chosen for displaying sans serif text characters in text views. Set in the Display Options dialog box. Compare with **serif font mapping** and **monospaced font mapping**.
scanner	A peripheral device that can convert (or digitize) the image of a page into digital form for use by a computer. A scanner is similar to a photocopier, but instead of producing a hard copy result on paper it sends its results electronically over a cable hooked up to a computer.
scanner driver	The system file that identifies a scanner to the system. It typically contains the I/O address of the scanner and specific information about the scanner's characteristics.
scanning	The act of using a scanner to convert (or digitize) the image of a page into digital form for use by a computer.
screen font	The representation of a font or typeface used for display on a screen. See also **font**, **font mapping**, and **typeface**. Compare with **printer font**.
scroll bars	A Pro OCR window contains two scroll bars—the vertical scroll bar and the horizontal scroll bar—that enable you to move around on a page beyond the screen boundaries, when necessary.
serif	The small decorative stroke at the ends of characters in some typefaces. Also, the designation for font families in which the characters have serifs. Common serif font families include Times Roman, Palatino, and Garamond. Compare with **sans serif**.
serif font mapping	The font chosen for displaying serif text characters in text views. Set in the Display Preferences dialog box. Compare with **sans serif font mapping** and **monospaced font mapping**.

settings file	A file, saved by choosing Save Settings from the File menu, that saves the current gallery, processing preferences, display preferences, proofing preferences, and selected scanner information in a named settings file. To use a settings file, retrieve it by choosing Retrieve Settings from the File menu.
sheetfed scanner	Scanner with an integral sheetfeeder, but no flatbed, on which pages are placed and fed through the scanner. Although they can scan multiple pages at a time, sheetfed scanners often support only a small range of paper sizes and it's difficult to control the proper alignment of a page. Compare with **flatbed scanner** and **automatic document feeder (ADF)**.
side-by-side columns	Also known as "bound" columns. A column format where the text flows as in a table, left to right, by column groups. Side-by-side columns are commonly found in tables and documents where the text reads left-to-right, then top to bottom. The flow of text in side-by-side columns is best suited for the As Single Column locate setting in Pro OCR.
single-bit image	Also referred to as line art. An image format where individual pixels are expressed as a single bit—either black or white. Compare with **grayscale image**.
single-step processing	A method for using Pro OCR with maximum control over individual pages. Single-step processing involves selecting Gallery settings for individual pages in a document, and manually launching Get Page, Locate and Recognize. Single-step processing is best suited to documents that require different Gallery settings (Page Size, Brightness, Locate method, etc.) on different pages. Compare with **automatic processing**.
skewed text	Text that is not horizontal in the page image. The most common cause of skewed text is scanning a page in crooked. Sometimes, text may be skewed on the input page. Pro OCR can accurately recognize text skewed up to 2°. If text is skewed more than that, Pro OCR may have difficulty in properly locating text regions. Problems with skewed pages (up to 15°) can be eliminated by selecting the Straighten Skewed Images setting in the Processing Options dialog box.

spell checking	Pro OCR automatically checks spelling during the Recognize step using its built-in dictionary and the current user dictionary. After Pro OCR finishes recognizing, you can check the spelling in a document using the user-configured Proof command.
standard resolution	A term associated with FAX modems, referring to the default resolution of the image files produced by these devices. Standard resolution is approximately 200 x 100 dpi, which may be insufficient for reliable recognition.
status bar	The panel of controls located along the bottom edge of the Pro OCR window. The status bar contains the view size selector, page indicator, view selector, and status display area.
status display area	At the right end of the status bar. The status display area shows the percentage of the current process that is completed. After recognition this area shows the number of suspect and illegible characters in the current page.
Stringent suspect threshold	Tells Pro OCR to highlight all suspect characters. Use it when accuracy is important and when there are many words in the document that are not in the dictionaries. Set in the Display Options dialog box. Compare with **Lenient suspect threshold** and **Normal suspect threshold**.
stroke weight	A measure of the average distance between the edges of the lines in a character. Certain typefaces have heavier stroke weights than others. A bold typeface has a heavier stroke weight than a Roman typeface.
Style ribbon	The panel located just beneath the Gallery inside of the Pro OCR window. The Style ribbon makes it quicker and easier to find and choose various style attributes for locate regions and selected text. See also **region style** and **text style**.
stylized font	A font with exaggerated serifs and embellishments and/or extraneous lines. Stylized fonts are a problem for the so-called omnifont (feature extraction) systems because these fonts do not adhere to generic character format rules required by omnifont technology. Zapf Chancery is a stylized font.

subscript text	Text with the _{subscript attribute} is below the baseline like this.
superscript text	Text with the ^{superscript attribute} is above the baseline like this.
supplementary dictionaries	Optional dictionaries that can be used during spell checking in Pro OCR. There are four supplementary dictionaries included with Pro OCR: geographical, legal, medical, and an expanded dictionary. Compare with **built-in dictionary** and **user dictionary**.
suspect character	A character that Pro OCR recognized with less than total confidence. Suspect characters in a document are highlighted in the text view. Compare with **illegible character**. See also **suspect character**.
suspect threshold	Pro OCR has three thresholds for highlighting suspect characters: Stringent, Normal, and Lenient. Each suspect character has a confidence value associated with it. Setting the suspect threshold determines the minimum confidence value used to highlight suspect characters. A lenient threshold displays only the suspect characters with the lowest confidence values, while a stringent threshold displays all suspect characters.
Tag Image File Format	See **TIFF**.
template	A previously saved file that defines and applies the locate regions on the pages of a document.
template matching	An older OCR technology where the application is trained by the user to recognize certain fonts by providing whole-character samples to be referenced against an unknown character until a suitable match is found. In practice, limited to recognizing a few specific fonts (typeface and point size).
Template locating method	One of Pro OCR's three locating methods. Use it to specify preset locate regions. Compare with **Normal locating method** and **As Single Column locating method**.
text quality	See **type quality**.

text region	Defines a text area on the page image in the image view and the text view. Only text within defined text regions is recognized. Text regions may be defined manually or by using Pro OCR's automatic locating settings.
text style	A piece of text's attributes or styling, such as **bold**, *italic*, or underline. Use the Style menu or the style ribbon to set these attributes. See also **bold text**, **italic text**, **underline text**, and **Style ribbon**.
text view	The view that displays the recognized text from the page image. You can proof and edit recognized text in the text view.
throughput	A measure of the total time required to reproduce printed documents. This effort measurement accounts for scanning time, recognition accuracy, error correction and format retention. Throughput is a more illuminating measure of OCR effectiveness than the simplistic recognition accuracy criterion commonly used to evaluate OCR performance. Compare with **recognition accuracy**.
TIFF	(Tag Image File Format) Standard graphic file format for saving high-resolution bitmapped images. Pro OCR can read most single-bit TIFF files produced by many scanners and applications. Pro OCR also saves to its own proprietary TIFF format. See also **Pro OCR format**.
touching characters	Character elements of an image where the spacing of the characters is insufficient to easily determine proper character boundaries. For example, in a document with touching characters, it may be difficult to differentiate between the letter pair "rn" and the character "m." Problems with touching characters can be reduced by using the brightness setting in Pro OCR to lighten the image.
typeface	One style within a font family. For example, Helvetica Bold Italic is a typeface. See also **font** and **font family**.
type quality	A quality of printed matter. Pro OCR offers a choice between Letter Quality or Draft Quality. See also **letter quality text** and **draft quality text**.

type size	The vertical height measurement of type, commonly expressed in points (72 points=1 inch). Pro OCR recognizes and preserves type ranging in size from 5 points to 64 points.
type style	The variations in characters, including font characteristics such as bold and italic, and styling characteristics such as underlining. Pro OCR recognizes and preserves many type style characteristics.
underline text	Text with the underline attribute looks like this. See also **text style**.
User Defined page size	One of Pro OCR's page size options when scanning. You may set the page size from 1" x 1" up to the limits of your scanner.
user dictionary	A dictionary file that the user may add words to. It is used along with the built-in dictionary to assist in recognition and to mark possible misspelled words. Compare with **built-in dictionary** and **supplementary dictionaries**.
view selector	The second set of controls from the left in the Status bar. Use it to quickly change between the image view and the text view. One of the two view icons is highlighted to indicate which view you're currently in.
window	An area that displays information on a desktop; you view a document through a window. You can open or close a window, move it around on the desktop, and sometimes change its size, scroll through it, and edit its contents.
Windows	The application interface manufactured by Microsoft Corporation that provides a graphical user interface (GUI) based upon a desktop, windows, menus, and icons.

word wrap	The automatic continuation of text from the end of one line to the beginning of the next. Word wrap lets you avoid pressing the Return key at the end of each line as you type. For example, when you input text in most word processors, lines of type are automatically “wrapped” to the next line when they won’t fit within the current line margins. If you change the margins, or the type size, or the spacing between words in a document, lines are often re-wrapped. When you save documents in any export format, text lines are wrapped in the output file. When you save documents in ASCII format, you can prevent lines from wrapping and preserve specific line breaks by selecting the option to preserve line breaks in the Save As Options dialog box.
zoom controls	The first set of controls at the left end of the status bar. Use them to easily change between magnification (zoom) levels.

Contents

Chapter 2: Learning Pro OCR Basics

The Basic Steps

Starting Pro OCR

Selecting a TWAIN-Compliant Scanner

Learning About the Gallery Toolbar

Tutorial Examples

Example 1: Using Auto OCR to Scan a One-Page Simple Document and Save It in Pro OCR Format

Example 2: Opening a File and Saving It in a Word Processor Format

Example 3: Scanning a Document of Multi-Column Text

Example 4: Scanning a Document With Tables and Saving in a Spreadsheet Format

Example 5: Scanning and Saving a Document with Pictures

Example 6: Locating a Document Using a Template

Example 7: Scanning a Document with Mixed Tables and Manually Locating Regions

Glossary

Chapter 2

Learning Pro OCR Basics

This chapter gets you started with Pro OCR. It introduces you to the Pro OCR window features, tells you the basic steps that you use when you work with Pro OCR, and provides several tutorial examples that you can complete to practice with Pro OCR.

TIP: If you use PaperPort software or scanners, see the *Working with PaperPort* document that came with Pro OCR. It provides tips and other information about using Pro OCR with these Visioneer products.

The Basic Steps

When you use Pro OCR, you convert an image of text and save it an editable format. To complete this conversion you perform the following basic steps:

1. **Get Page—**acquire pages either from a scanner or by opening an image file.

2. **Locate—**indicate which text on the page you want to recognize, and which pictures (if any) to retain.

3. **Recognize—**convert the image to text.

4. **Proof—**check for incorrectly identified and unidentifiable characters and make changes to recognized text.

5. **Save**—save the text to a variety of application formats.

Often, you automatically complete the first three steps by clicking the Auto OCR button, however, you can perform each step individually. You can also use a combination of automatic and individual processing by using deferred and finish processing features.

Starting Pro OCR

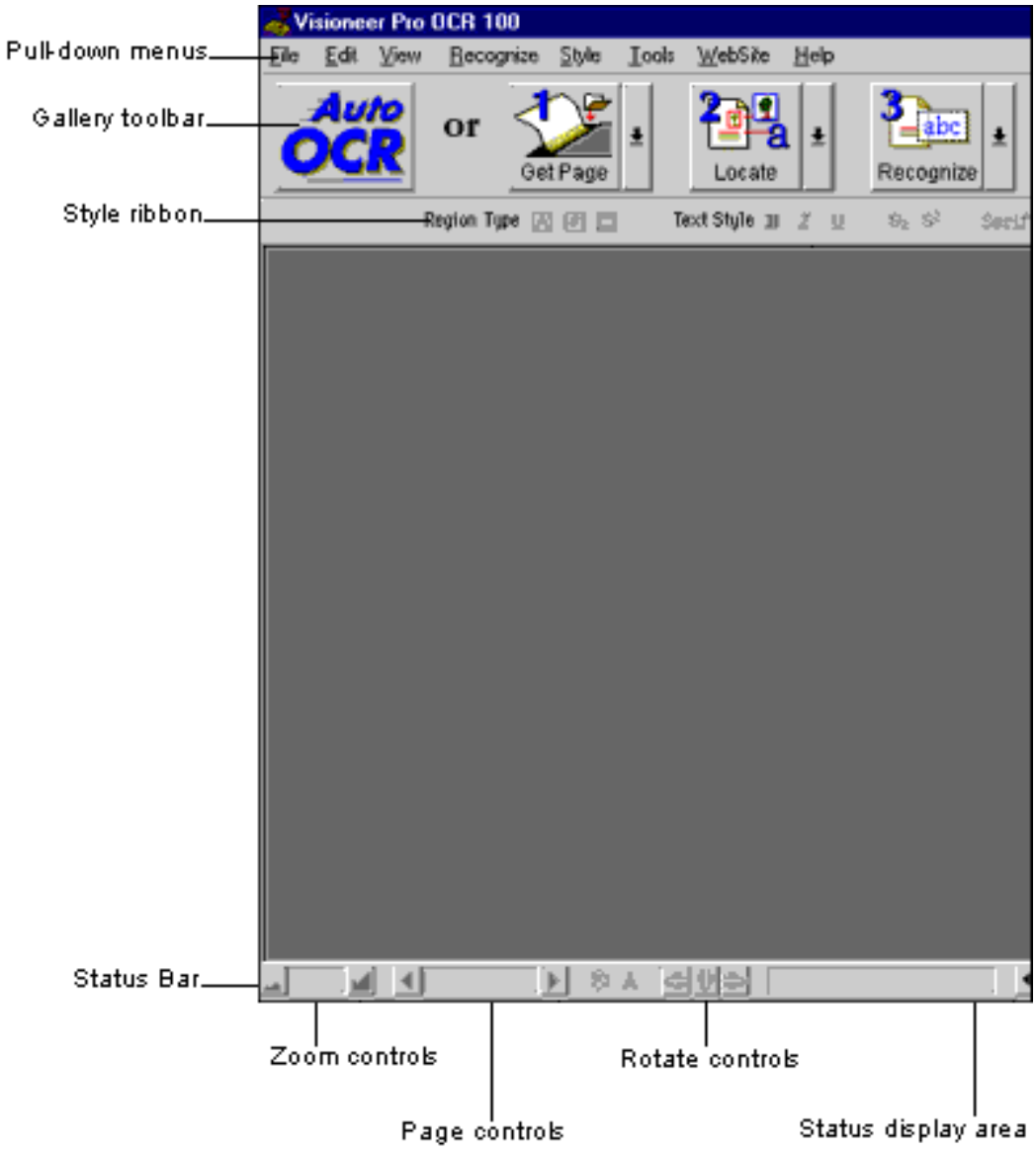
The following procedure helps you to get acquainted with Pro OCR and make sure that everything is set up correctly.

TIP: In addition to the following procedure, Visioneer provides two other ways to start and use Pro OCR: 1) From the Windows Start menu, choose Programs, and then choose Visioneer OCR Wizard. 2) If you use PaperPort software, start PaperPort and then choose the Pro OCR link.

To start Pro OCR and select processing options:

1. From the Windows **Start** menu, choose **Programs**, and then choose **Visioneer**. From the **Visioneer** menu, choose **Visioneer Pro OCR 100**.

   The Pro OCR window appears. It includes pull-down menus, the Gallery toolbar, the Style ribbon, and the Status bar.



Feature	Does this...
Pull-down menus	Contains commands and options that you use to set process options and initiate actions. Many of the commands in the pull-down menus are also available by using the Gallery buttons and Gallery buttons drop-down lists.
Gallery toolbar	Lets you change common settings, start Auto OCR, or individually perform any of the basic steps required to convert an image to text. Several Gallery buttons have drop-down lists from which you can select options.
Style bar	Makes it easy to choose various style attributes for selected regions and text. The Region Type options are available in image view and the Text Style options are available in text view.

Status bar	Contains controls with which you choose how to view pages (text or image view) and which pages to view. The Status bar also contains a status display area to keep you informed of Pro OCR's progress.
Zoom controls	Magnifies or reduces the view of the document.
View controls	Displays the page in a landscape or portrait orientation.
Page controls	Displays the previous or next page.
Suspects or Illegibles	Displays the number of suspect or illegible characters in the document.

Selecting a TWAIN-Compliant Scanner

Before you scan an item with Pro OCR, make sure the scanner software is installed, and the scanner can scan images into your computer. Pro OCR works with many TWAIN-compliant devices. You can select the TWAIN device in the Pro OCR software.

NOTE: If you are using Pro OCR with Visioneer's PaperPort software or scanners, see the *Working with PaperPort* document that came with Pro OCR, instead of the following procedure. If you are using a scanner that is not TWAIN-compliant, you cannot scan directly to Pro OCR. Instead, use your scanner's software to save the scanned file in a TIF format, and then use the Pro OCR Get File command. For more information, see "Getting Pages from an Image File" in Chapter 3.

To select a scanner:

1. Choose **Select Scanner** from the **Tools** menu.

   The Select Source dialog box appears.

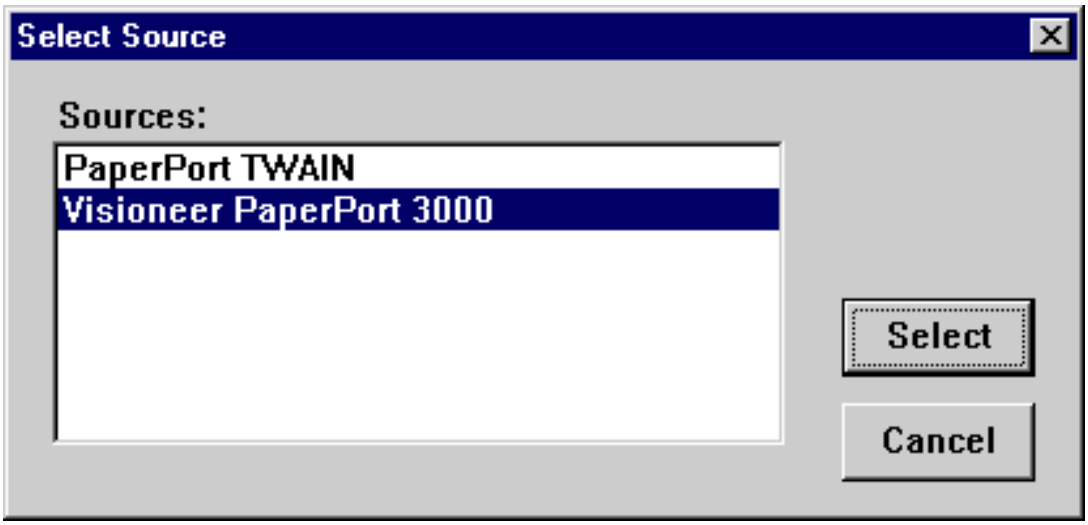


Figure 2-1: Select Source Dialog Box

NOTE: If the scanner driver you want is not shown, make sure that the scanner is properly connected to the computer and that both the scanner and the computer are plugged in, turned on, and operating correctly.

2. In the **Select Source** dialog box, select the TWAIN scanner driver you want to use with Pro OCR.

3. Click **Select**.

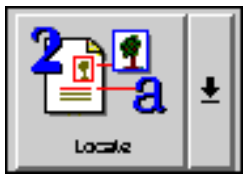
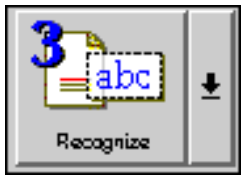
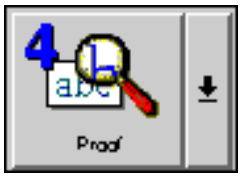
   The scanner you selected is available until you select a different one. You don't have to repeat this procedure unless you want to select a different scanner.

Learning About the Gallery Toolbar

The Gallery toolbar contains buttons for starting the various steps of the Pro OCR process, including the Auto OCR button. The buttons numbered one through four are also important because you can select different options from drop-down lists before processing a document. For example, you can tell Pro OCR whether the document is one column or multiple columns. The options you select from these buttons affect the way that Auto OCR processes a document.

Click this button to have Pro OCR perform the first three steps of the process

NOTE: Often you will use Auto OCR to complete processing. However, sometimes it is better to perform each step individually. (This is also referred to as manual or single-step operation.) For example, you use the single-step procedures when you want to manually define locate regions, create a template, redo a step, recognize different type quality settings, or scan pages that have mixed orientations (portrait and landscape.)

	Button	Does this...
	Auto OCR	Performs Steps 1, 2, and 3 (Get, Locate, and Recognize) of the OCR process. Before you click this button, select processing options from the Get, Locate, and Recognize drop-down lists.
	Get Page	Scans a page or opens an image file.
	Locate	Locates areas of text, pictures, and numbers and determines how text flows on the page.
	Recognize	Converts areas of the page into editable text.
	Proof	Checks the document for errors.

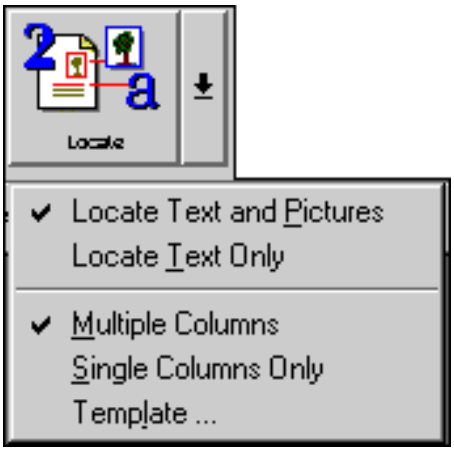
	Save As	Saves the converted document in a variety of formats, such as text, Rich Text Format (RTF), or HTML.

You can select options with the Gallery buttons by using the drop-down list next to each button.

To select an option from a Gallery drop-down list:

1. Click the arrow next to the Gallery button you want.

   The drop-down list for the button appears. The following figure shows the Locate button with the drop-down list displayed.



2. Select the option you want.

   A checkmark appears next to the option you selected.

Tutorial Examples

Now that you know the basic steps you can practice them using the sample documents that came with Pro OCR. The Pro OCR software comes configured and ready to use so that you don't have to change the various options. You can find copies of the pages that you scan for the tutorials in the back of the *Getting Started Guide*. You can also find sample files in the Pro OCR directory.

NOTE: If you don't have a scanner, you can complete the following exercises that require scanning, by instead using the Get File command and selecting the file from the Pro OCR directory.

Example 1: Using Auto OCR to Scan a One-Page Simple Document and Save It in Pro OCR Format

This example shows how to convert (recognize) the text in a one-page document. You can find a ready-to-use sample in the back of the *Getting Started Guide*.

Selecting Gallery Options

Pro OCR processes a document using the options that are set in each drop-down list associated with a button of the Gallery toolbar.

To set Gallery options for this example:

1. From the **Get Page** drop-down list, choose **Use Scanner**.

2. From the **Locate** drop-down list, choose **Locate Text Only** and **Single Columns Only**.

3. From the **Recognize** drop-down list, choose **Degraded or Fax Quality**.

Starting Auto OCR

By clicking the Auto OCR button, you can perform the first three steps of the OCR process, that is, Get Page, Locate, and Recognize.

To process a simple document without any graphics:

1. Remove Sample A from the back of the *Getting Started Guide*. The document is a simple business letter.

2. Place the document on the scanner.

3. Click the **Auto OCR** button.

   When you click Auto OCR, your scanner software dialog box appears.

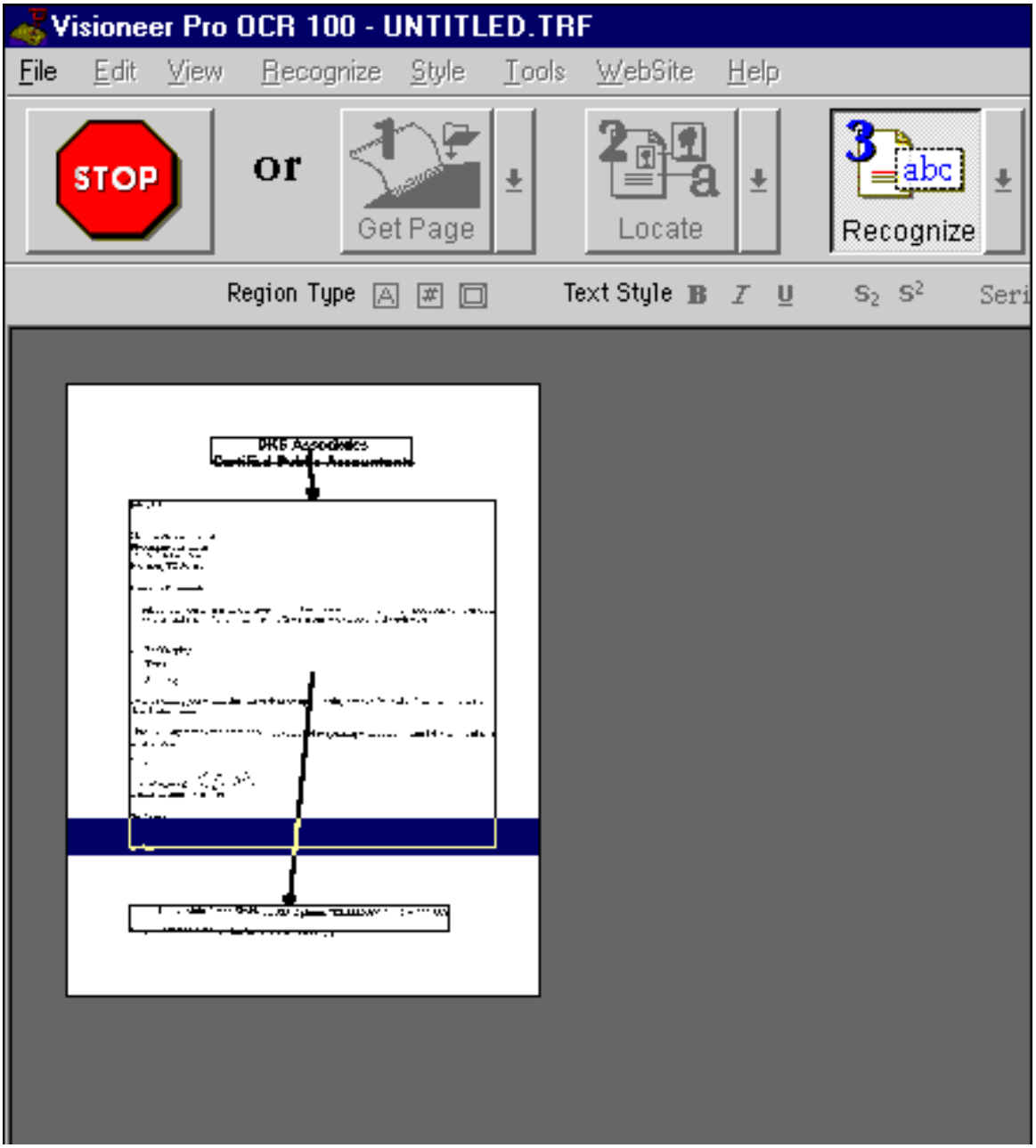
4. Use the scanner software as you usually do to scan a page.

5. After the scanner has scanned the page, Pro OCR displays a dialog box that asks if you want to scan another page.

6. Click **End**.

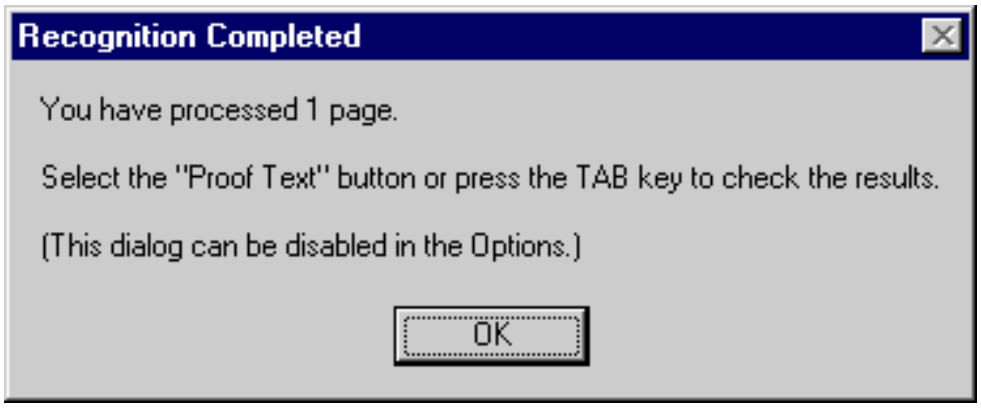
Pro OCR continues with the second task to locate text regions on the page.

A progress bar moves down the page. When Pro OCR finishes locating, it displays text boxes indicating located text regions, with arrows connecting each text region to the next. Pro OCR outputs text in the order in which the arrows connect the text regions.



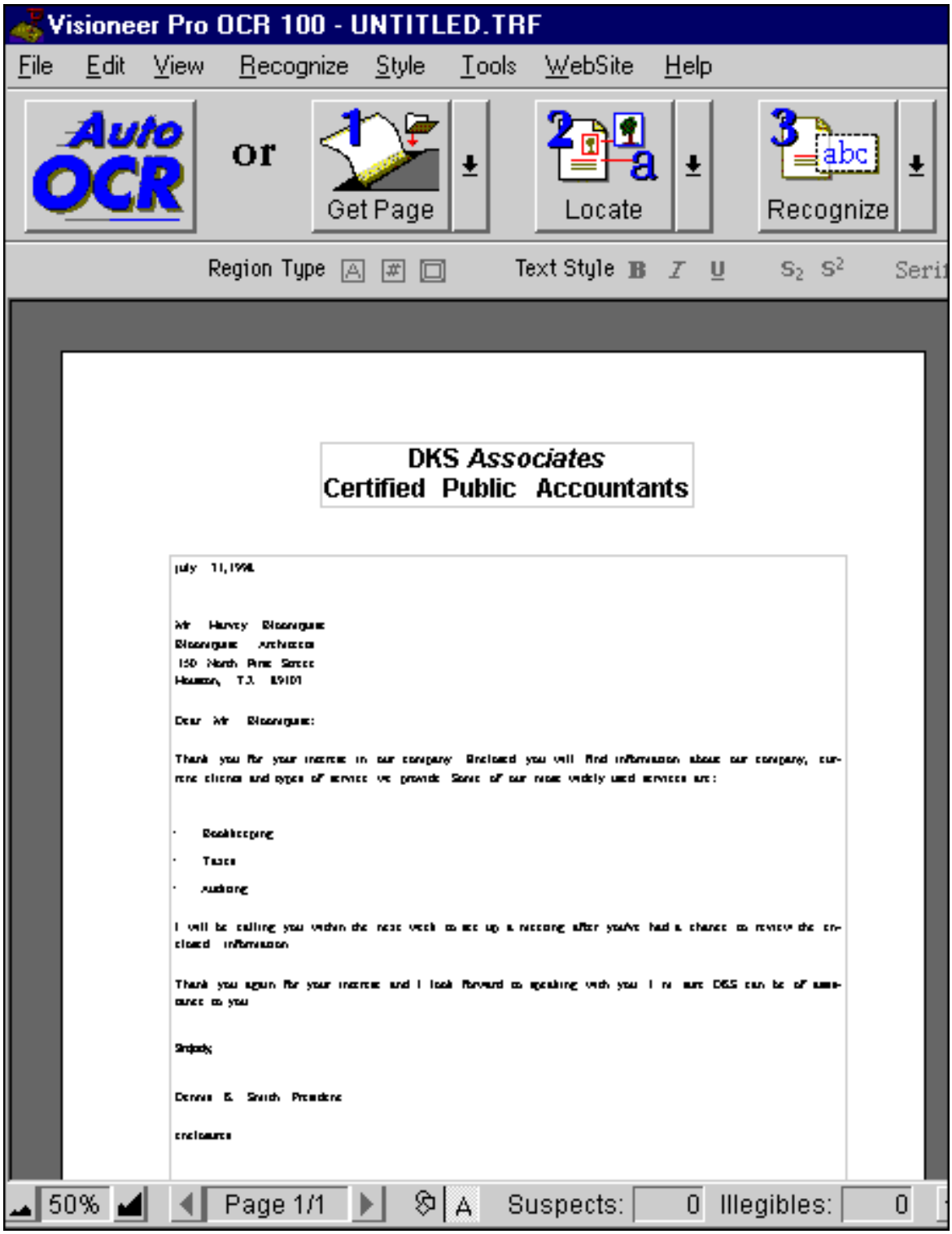
In the next step, Pro OCR recognizes the located text. While Pro OCR is recognizing, again a progress bar moves down the page.

When Pro OCR finishes recognizing the text, the Recognition Completed dialog box appears.


Recognition Completed
You have processed 1 page.
Select the "Proof Text" button or press the TAB key to check the results.
(This dialog can be disabled in the Options.)
OK

7. Click **OK**.

   The document appears in the text view. You use the text view to proof the document and correct any errors.



Usually at this point you proof the document. For now, just save it.

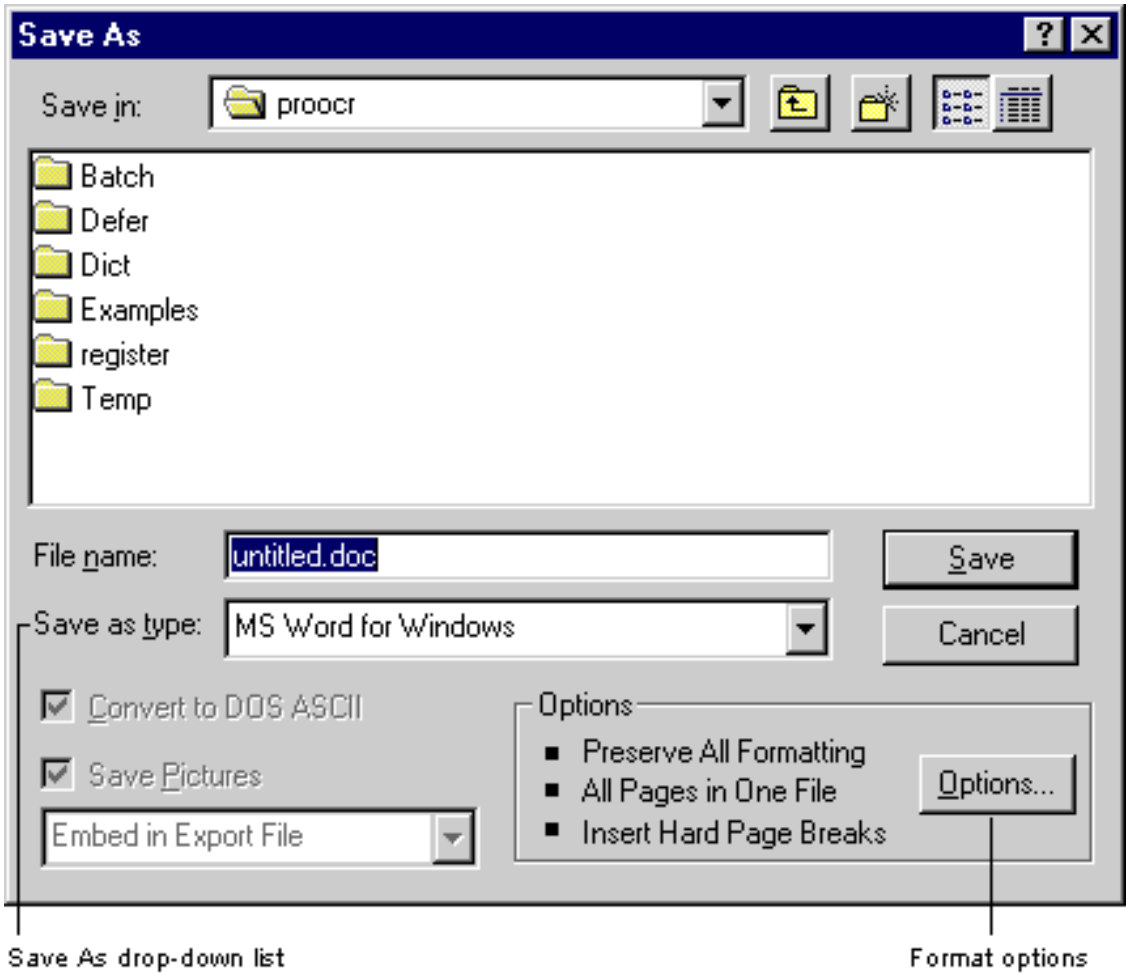
Saving a Document

You can save the processed document to disk in different formats. For example, if you want to open the document again in Pro OCR, you select the Pro OCR format.

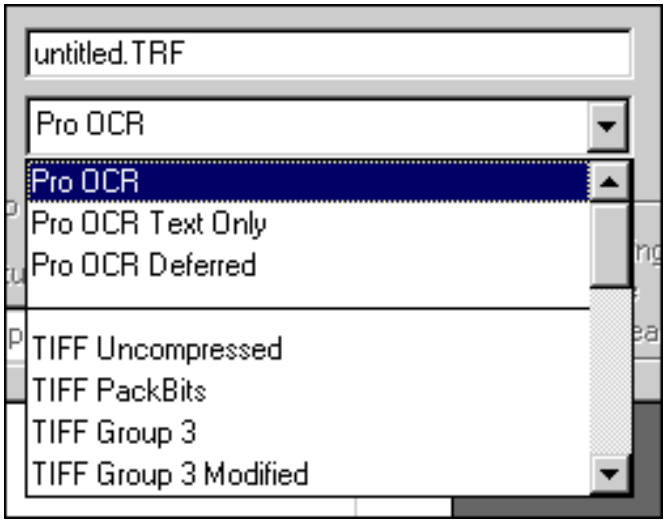
To save the document:

1. Choose **Save** from the **File** menu, or click the **Save As** button on the Gallery toolbar.

   The Save As dialog box appears.



2. Choose **Pro OCR** from the **Save As** drop-down list.



   By saving the document in this format, you can edit the pages later within Pro

OCR. If you save in another file format, you must open it in an application that supports that format.

3. Type in a name for the file in the **File Name** box.

4. Click **Save**.

   The text and format information of the document is saved in the format you've selected.

5. Choose **Close** from the **File** menu.

   You just completed your first job using Pro OCR. Many of the jobs for which you use Pro OCR are as quick and simple as this one. You can now continue by completing the rest of the examples in this guide.

Example 2: Opening a File and Saving It in a Word Processor Format

Instead of getting and processing a document from a scanner, you can also process a file that was saved on disk. You can use this procedure to read TIFF, PCX, or DCX files produced by Pro OCR or other applications.

Opening a File

For this example, use the file, SAMPLEB.TIF, in the Pro OCR directory. This is a document that has a graphic. Because of the difference between this document and the one used in the previous example, you will change the options in the Gallery toolbar. Although this document has a graphic, let's assume you don't want to save the graphic.

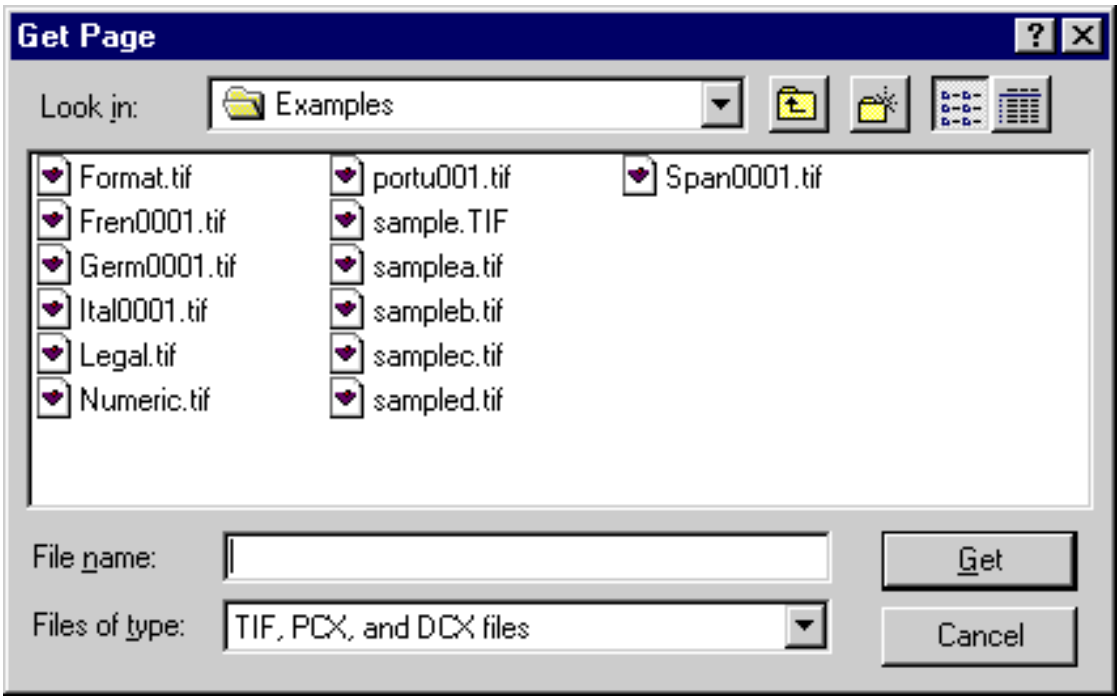
You can either set the options before each step or set them all at once. In this example, you'll set them as you go along.

To set the OCR options and get a file from disk:

1. Select **Open File** from the **Get Page** drop-down list.

2. Click the **Get Page** button in the Gallery toolbar.

   The Get Page dialog box appears.



3. In the Pro OCR directory, select the file **SAMPLEB.TIF**.

4. Click **Get**.

   The sample file is read in and the progress bar moves down the page.

Locating the Regions in a Document

For Pro OCR to properly convert areas of a document, you must locate the regions of the page that will be recognized. There are three types of regions: text, numeric, and picture. For example, a picture region is one that contains any kind of graphic, illustration, photograph, drawing, or picture. The contents of a picture region cannot be recognized, but can be saved as an image. By specifying the Locate options, Pro OCR knows what types of regions are in the document.

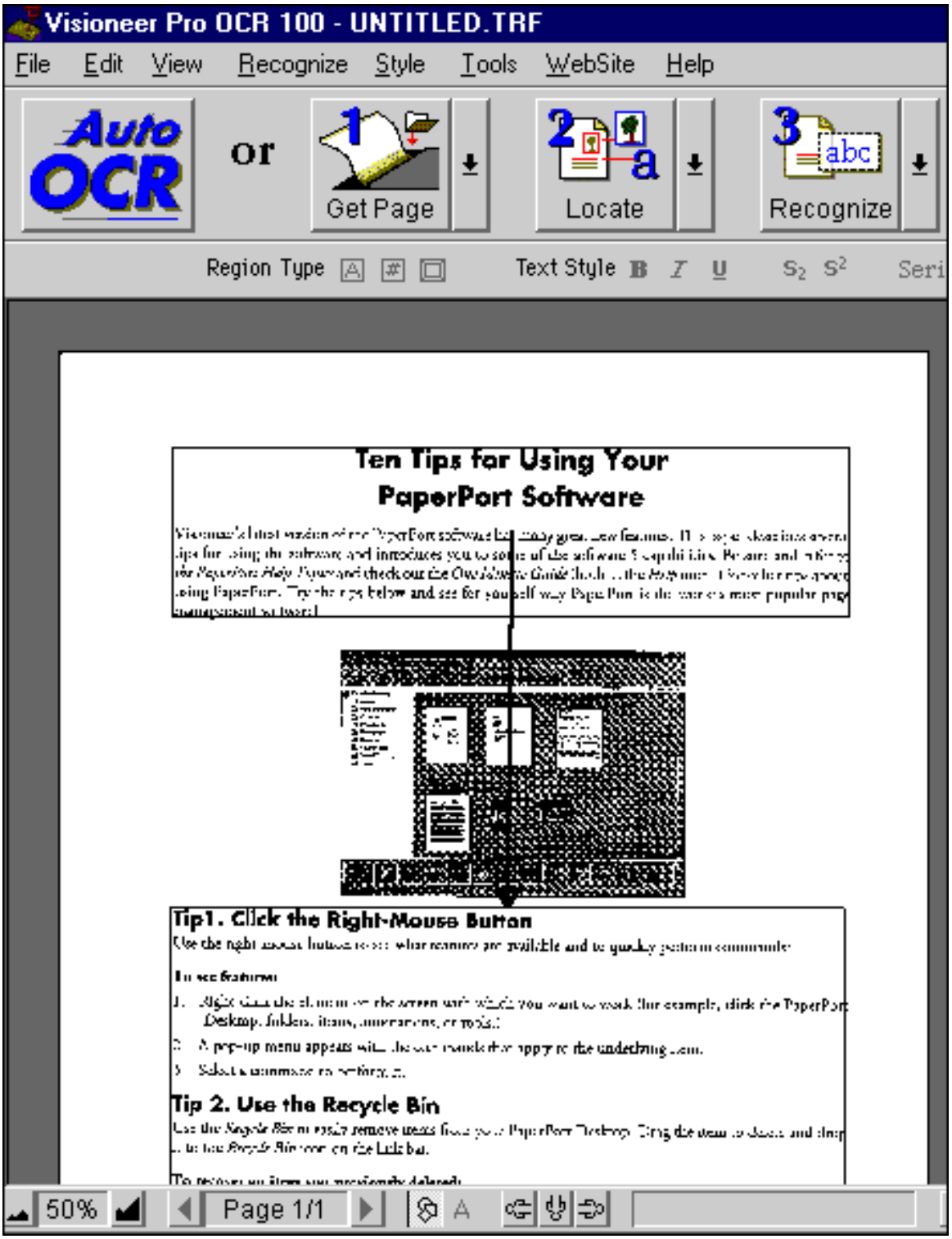
To specify the regions to locate:

1. Select **Locate Text Only** and **Single Column** from the **Locate** drop-down list.

   If you did want to save the graphics in a document, you would select Locate Text and Pictures. Sometimes, you want the graphics so that you can recreate an exact duplicate of the document you are processing.

2. Click the **Locate** button in the Gallery toolbar.

   Pro OCR goes through the document and recognizes the different regions. Arrows appear on the document showing the flow of the information.



Recognizing the Document

The third step is to actually convert or recognize the text in a document. Pro OCR reads the text and displays the actual characters.

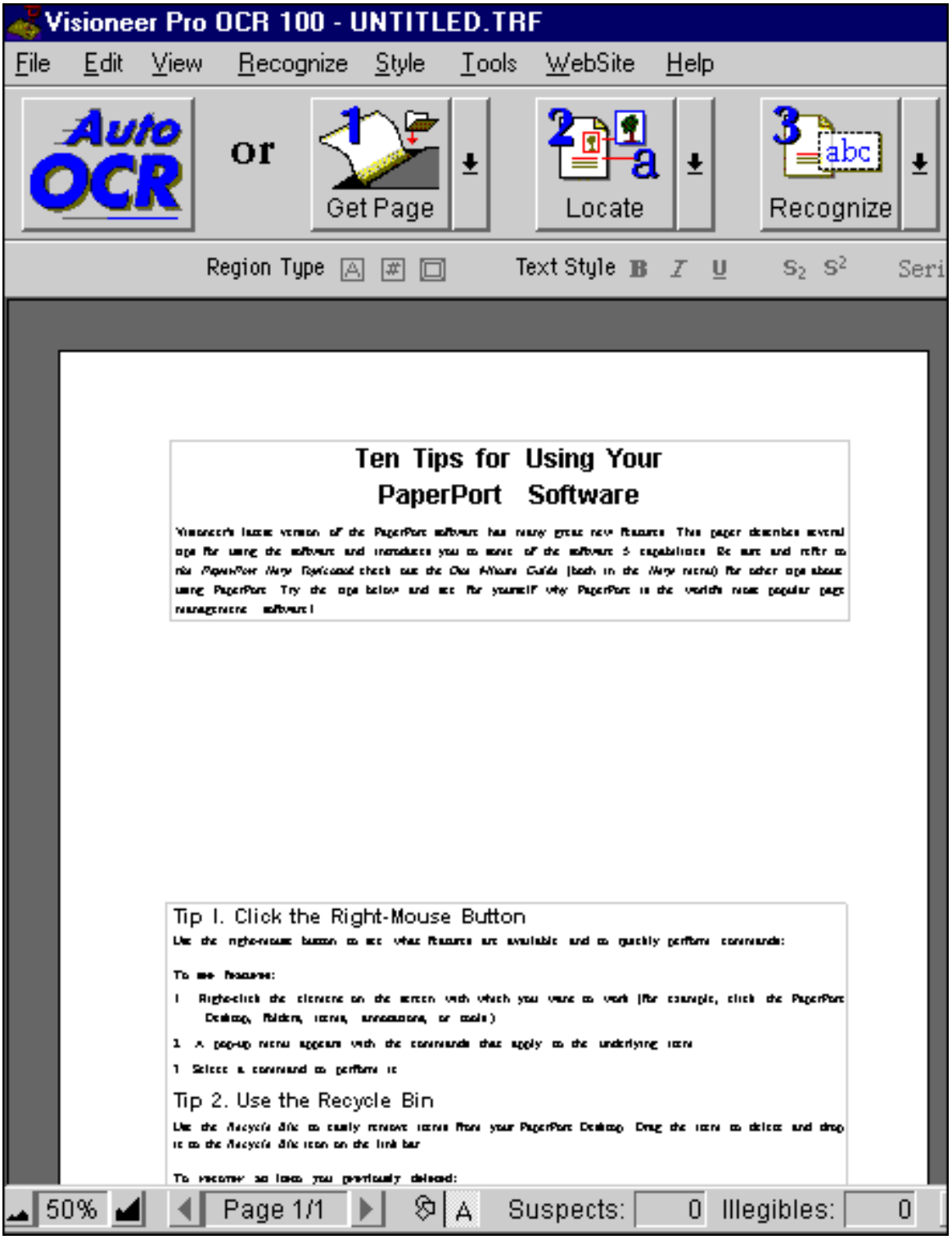
Before recognizing the document, you should specify the quality of the image text. You can do this by using the Recognize drop-down list.

To recognize the document:

1. Select **Degraded or Fax Quality** from the **Recognize** button drop-down list.

2. Click the **Recognize** button in the Gallery toolbar.

   Pro OCR displays a bar that moves through the document as Pro OCR recognizes the text. When the process finishes, you see the document with text only.



Proofing the Document

After a document is recognized it appears in the text view. In this view, you can proof the document for errors and make changes to the document when you find problems.

When you proof, you can:

- Inspect recognized text and edit it if necessary.
- Search for misspelled words, numbers, punctuation, symbols, and alphanumeric words.
- Change font style information.

NOTE: You can change the proofing options by choosing Options from the Tools menu.

To proof the document:

1. Click the **Proof** button in the Gallery toolbar, or press the **Tab** key.

   Pro OCR starts at the current insertion point, if there is one. Otherwise, it starts at the top of the current page.

   Pro OCR highlights the first word it does not recognize and displays the suspect text in the On-Screen Verifier.

   The On-Screen Verifier is a pop-up window that displays the part of the page image corresponding to selected text.

   TIP: For a a close up of the text, click the image to increase the magnification.

2. If the text is wrong, select the text and type the correct text.

3. Click the **Proof** button in the Gallery toolbar again or press the **Tab** key.

Pro OCR displays the next suspect entry.

4. Repeat the previous steps until you have checked the entire document.

5. If you want to change the font style, select the text, and click the Style option.

Saving the Document

Saving the document places a permanent copy of it on disk.

To save the document:

1. Choose **Save** from the **File** menu, or click the **Save As** button in the Gallery toolbar.

   The Save As dialog box appears.

2. Type the file name in the **File Name** box.

3. Select **MS Word for Windows** from the Save as drop-down list.

   You can save documents in many popular formats, including Rich Text Format (RTF), plain text, and Microsoft Excel.

4. Click **Save**.

5. Choose **Close** from the **File** menu.

Example 3: Scanning a Document of Multi-Column Text

This example introduces you to processing of multi-column text like newspapers, magazine articles, and multicolumn books (but not tables), where you want the text to be recognized column by column.

To scan multi-column text and save in Pro OCR format:

1. Put Sample Document C in the scanner. You can find a copy of this document in the back of the *Getting Started Guide*.

   Make sure to place the document in the correct orientation and to align it.

2. Select **Locate Text Only** and **Multiple Columns** from the **Locate** drop-down list in the Gallery toolbar.

Locate Text Only prevents Pro OCR from locating any picture element in the document to be scanned.

3. Select **Use Scanner** from the **Get Page** drop-down list in the Gallery toolbar.

4. Click **Auto OCR** in the Gallery toolbar.

   Your scanner software dialog box appears.

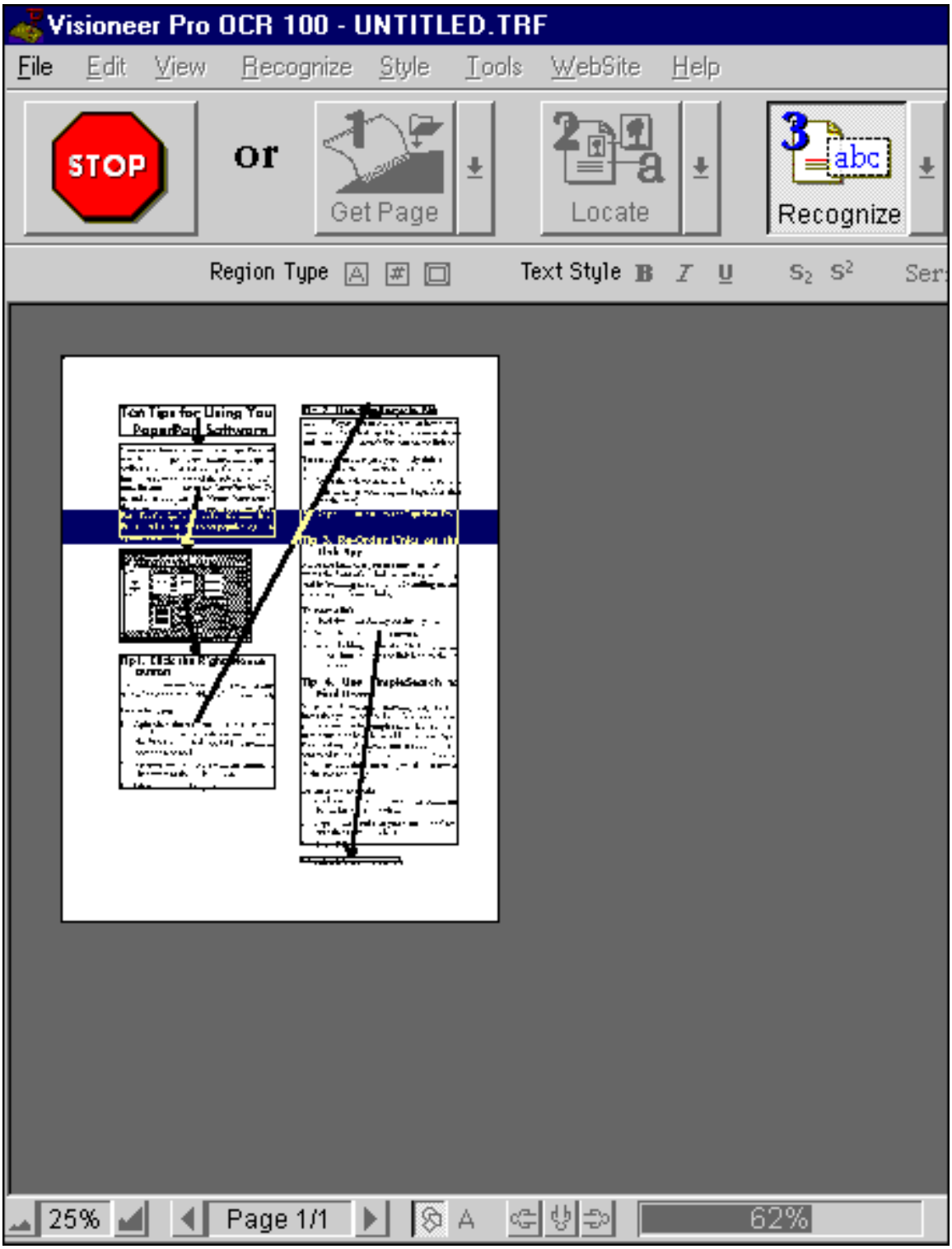
5. Use the scanner software as you usually do to scan the document.

   After scanning the sample document, the document appears in Pro OCR.

   A dialog box appears asking for additional pages to scan. For this example, you won't scan any additional pages.

6. Click **End**.

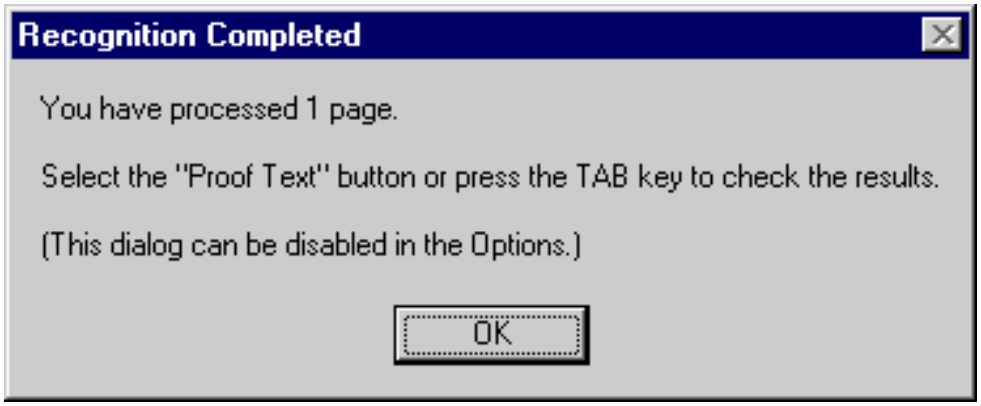
   Automatic processing continues with locating and then recognizing.



While Pro OCR recognizes the page, notice the boxes indicating located text regions around each column, and the arrows connecting each text region to the next. Note that by using Locate Text Only, the graphic element in the sample was not located and so a box does not appear around it.

Pro OCR outputs text in the order in which the arrows connect the text regions. For this example, notice how the boxes are drawn and connected.

When Pro OCR finishes recognizing, the Recognition Completed dialog box appears.


Recognition Completed
You have processed 1 page.
Select the "Proof Text" button or press the TAB key to check the results.
(This dialog can be disabled in the Options.)
OK

7. Click **OK**.

   The document appears in the text view.

To save the document

1. Choose **Save As** from the **File** menu, or click the **Save As** button in the Gallery toolbar.

   The Save As dialog box appears.

2. Select **Pro OCR** from the **Save As Type** drop-down list.

   The Pro OCR format saves all available information in the document.

3. Type in a name for the file in the **File Name** box.

4. Click **Save**.

   Both the image of the scanned page and the recognized text are saved. Always save files in the **Pro OCR** format when you want to reopen them in Pro OCR.

   NOTE: To reopen a file saved in the **Pro OCR** format, use the Open command from the File menu. If you use Get Page, Pro OCR only restores the page image. The Open command restores all the saved information, including any recognized text and proofing information.

5. Choose **Close** from the **File** menu.

   For information about other file formats, see Chapter 6, “Saving and Printing

Documents."

Example 4: Scanning a Document With Tables and Saving in a Spreadsheet Format

This example introduces you to processing of multi-column text in tables, where you want the text to be recognized as all one text block and not broken into columns. You can use this procedure whenever you want to recognize tables and other documents that you don't want broken into columns.

To scan multicolumn table text and save in spreadsheet format:

1. Select **Single Columns Only** and **Locate Text Only** from the **Locate** drop-down list in the Gallery toolbar.

2. Put **Sample Document D** in the scanner.

   Make sure to place it in the correct orientation to align it.

3. Click **Auto OCR**.

   Pro OCR displays your scanner software.

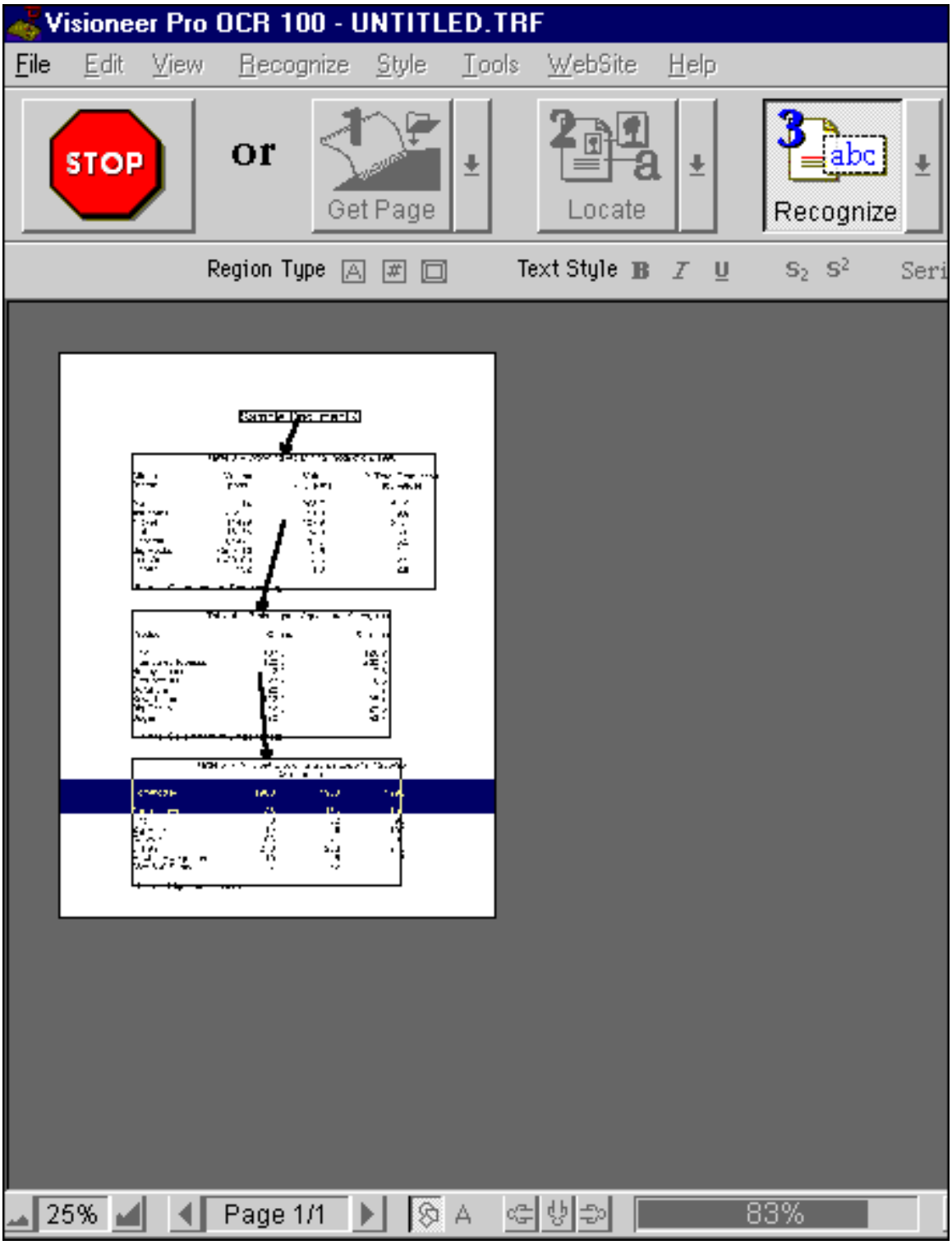
4. Use the scanner software as you usually do to scan the document.

   After scanning the sample document, it appears in the Pro OCR window.

   A dialog box appears. asking if you want to scan additional pages. For this example, you won't be scanning any additional pages.

5. Click **End**.

   Pro OCR locates and then recognizes the page.

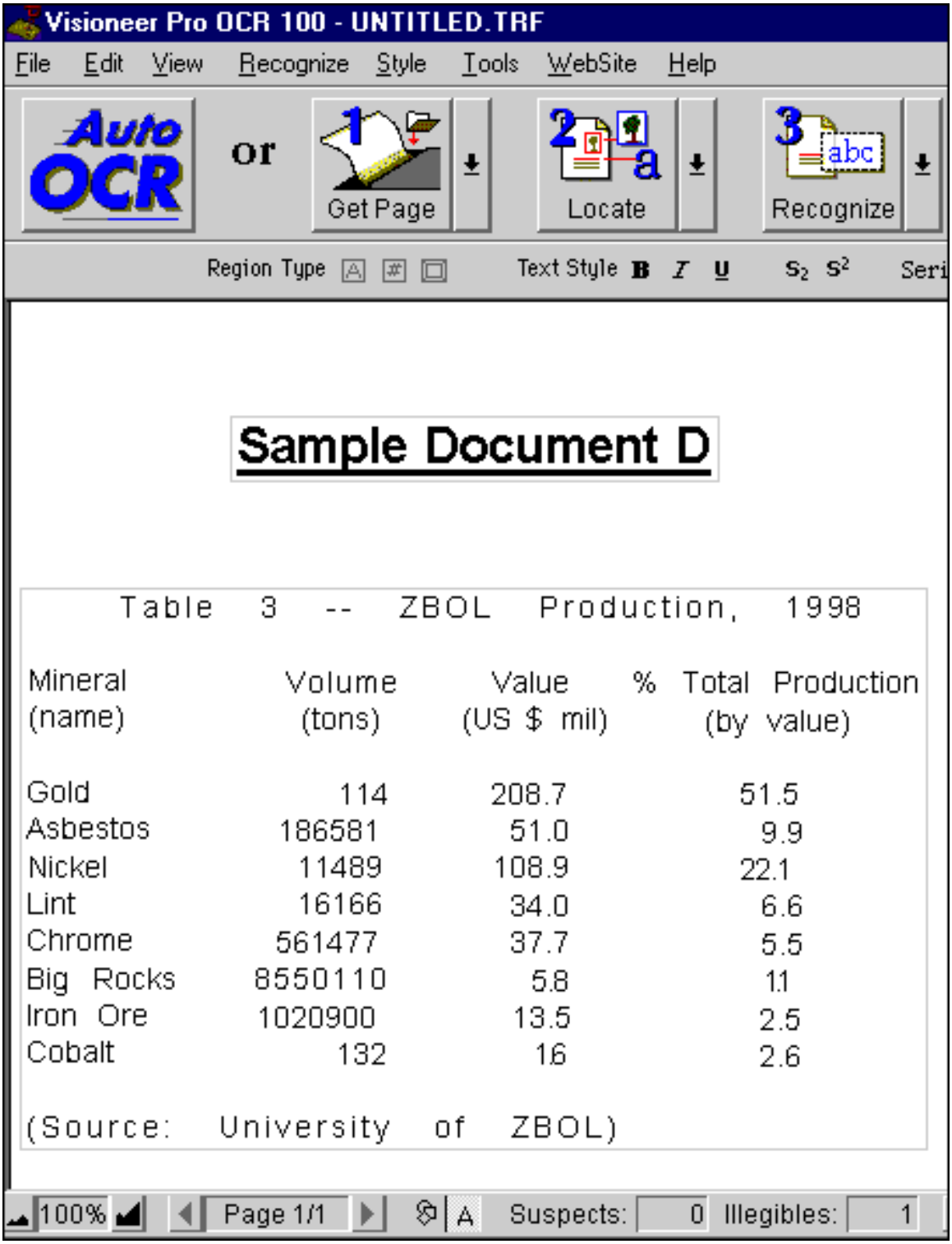


Notice that the text regions are not drawn separately around each column. By using the Single Column locating method, you force Pro OCR to ignore columns and tell it to read the page from left to right, top to bottom.

When Pro OCR is finished recognizing the page, the Recognition Completed dialog box appears.

6. Click **OK**.

Pro OCR displays the document in the text view.



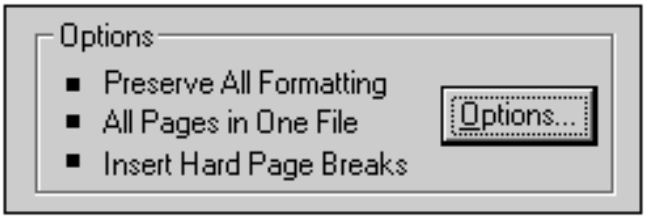
To save the document:

1. Choose **Save As** from the File menu, or click the **Save As** button in the Gallery toolbar.

   The Save As dialog box appears.

2. Choose **Microsoft Excel** from the **Save as Type** drop-down list.

   Notice that the following options are already selected.



   TIP: To change these options, click the Options button.

3. Type in a name for the file in the **File Name** box.

4. Click **Save**.

   Pro OCR saves the text and format information of the document in the format you have selected.

5. Choose **Close** from the **File** menu.

NOTE: If you don't save a version of this file in the **Pro OCR** format, you cannot open it again in Pro OCR. You can open the version that you just saved in any spreadsheet application that supports the Microsoft Excel format.

Example 5: Scanning and Saving a Document with Pictures

This example shows you how to scan a document with photographs or line drawings and save it in a word processor file format.

To scan and save a document with pictures:

1. Select **Multiple Columns** and **Locate Text and Pictures** from the **Locate** drop-down list in the Gallery toolbar.

2. Put Sample Document C in the scanner. You can find this document in the back of the *Getting Started Guide*.

3. Click **Auto OCR**.

   Pro OCR displays your scanner software.

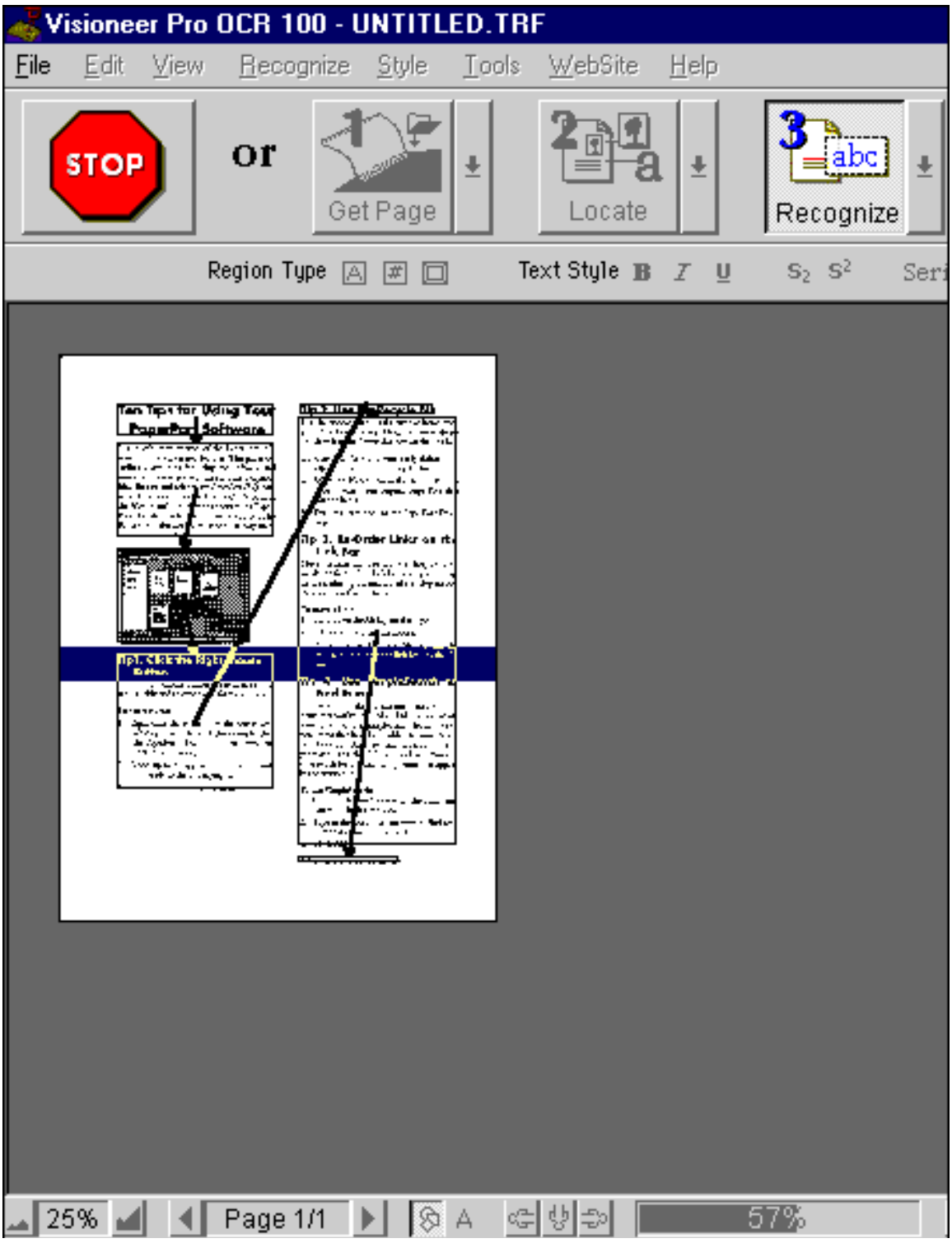
4. Use the scanner software as you usually do to scan a document.

After scanning the sample document, it appears in the Pro OCR window.

Pro OCR begins getting the page from the scanner. When the scanning is done, a dialog box appears asking if you want to scan additional pages. For this example, you won't be scanning any additional pages.

5. Click **End**.

Automatic processing continues with the Locate and Recognize steps.



The Recognition Complete dialog box appears.

6. Click **OK**.

   The document appears in the text view. Notice that the graphic image appears and has a picture region drawn around it.

To save the document:

1. Choose **Save As** from the **File** menu, or click the **Save As** button in the Gallery toolbar.

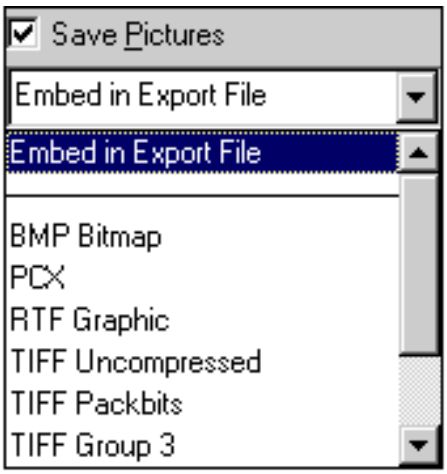
   The Save As dialog box appears.

2. Choose **Rich Text Format (RTF)** from the **Save as Type** drop-down list.

   RTF allows you to save the pictures along with the text in the exported file.

   NOTE: As an alternative, you can save in a format for an application that you have, such as Ami Pro, Word for Windows, and WordPerfect 5.x.

3. Select the **Save Pictures** option.



4. Choose **Embed in Export File** from the **Save Pictures** drop-down list.

   This format embeds the pictures into the RTF file along with the text.

5. Type a name for the file.

6. Click **Save**.

   The picture from the scanned page is now saved within the RTF file along with the recognized text. If you open this file in a word processor that supports pictures in RTF files, you see the recognized text and the pictures.

7. Choose **Close** from the **File** menu.

Example 6: Locating a Document Using a Template

At times, you don't want to recognize all the text on a page. For example, in this exercise the sample page has a header and a footer that you don't want to recognize or save. The sample template in this example is designed to create a text region around just the body text during the Locate step. The title and copyright in the footer are not recognized (saving time during recognition) and are not displayed (saving you the time of searching for and deleting them).

In this example, you use a supplied template that you can use for your own documents as well. You can also create your own templates, to customize Pro OCR for the kinds of pages that you typically use.

To use a template:

1. Choose **Template** from the **Locate** drop-down list in the Gallery toolbar.

2. Choose **Select Template** from the **File** menu.

   The Select Template dialog box appears.

3. In the **Temp** folder, find and select the file **TEMPB.TPL**.

4. Click **Open**.

   Pro OCR displays the name of the template you selected next to Template in the Locate drop-down list.

5. Select **Open File** from the **Get Page** drop-down list.

6. Click the **Get Page** button in the Gallery tool bar.

7. In the Pro OCR directory, select the file **SAMPLEB.TIF** and click the **Get** button.

   The sample file is read in.

8. Click the **Locate** button.

   Notice that text boxes are drawn around just the body text on the page. This is the text region defined by template. Only the text within this text region is recognized.

9. Click the **Recognize** button.

   After recognizing is completed the document appears in the text view. You can review the recognized document in the text view. Notice that the title and the

copyright in the footer were not recognized. If you save this page in an application or text format, only the displayed text is saved.

10. Save and close the document.

    Use the same procedures described in the earlier examples.

Example 7: Scanning a Document with Mixed Tables and Manually Locating Regions

This example shows you how to scan and manually locate a document with a table that has some rows or columns suitable for numeric regions and other rows or columns suitable for text regions.

To scan and locate a document with mixed tables:

1. Put Sample Document D in the scanner.

   Make sure to place it in the correct orientation and to align it.

2. Select **Single Column** from the **Locate** drop-down list.

3. Click the **Get Page** button.

   Pro OCR begins getting the page from the scanner and displays your scanner software.

4. Use your scanner software as you usually do.

   Pro OCR scans in the page and then displays it in the image view.

5. Choose **Zoom Out** from the **View** menu, or click the **Zoom Out** button on the Status bar.

   You can reduce or enlarge the document on the screen by using the Zoom In or Zoom Out features.

To select regions manually:

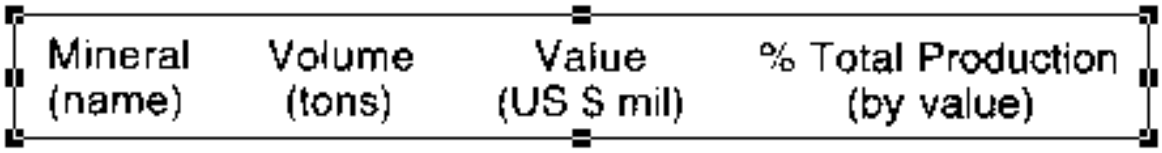
1. Scroll the page up a short distance so that the table labeled “ZBOL Mining Production, 1998” is fully visible on your screen.

2. Move the pointer just above and to the left of the first column header, titled “Mineral.”

3. Press and hold the mouse button; then drag down and to the right until the box following the pointer encloses all of the column headers.

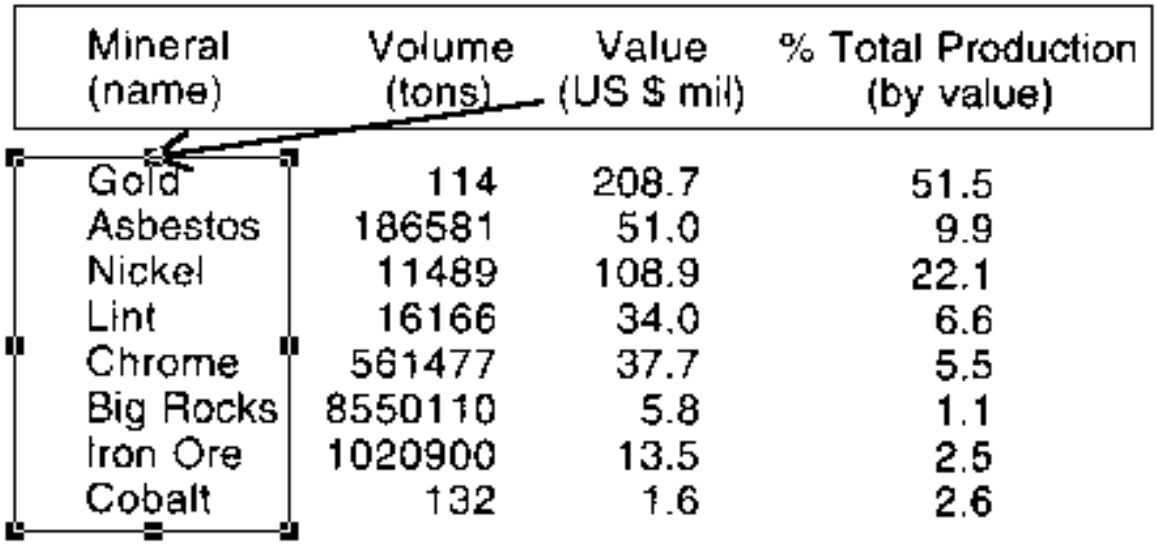
4. Release the mouse button.

   You have just manually located a text region.



5. Move the pointer just above and to the left of the item labeled “Gold.”

6. Press and hold the mouse button; then drag down and to the right until the box following the pointer encloses the first column of the table.



   The box should enclose the items from “Gold” through “Cobalt.”

   TIP: If you make a mistake, select the region and press Del.

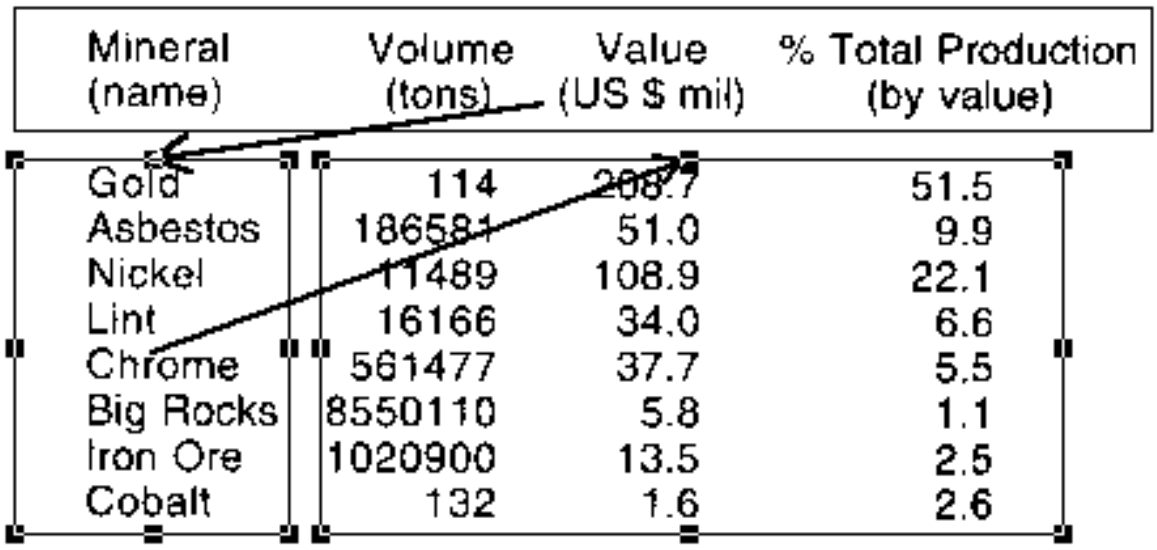
7. Release the mouse button.

   You have just manually located another text region. Note that an arrow appears that connects this text region to the first text region you defined for the table headers.

8. Move the pointer just above and to the left of the first number column.

9. Using the same steps you used to create the text regions, drag the mouse until the box following it encloses all three columns of numbers and release the mouse button.

   Make sure the entire image of the number columns is enclosed by the new region you have defined.



Mineral (name)	Volume (tons)	Value (US $ mil)	% Total Production (by value)
Gold	114	208.7	51.5
Asbestos	186581	51.0	9.9
Nickel	11489	108.9	22.1
Lint	16166	34.0	6.6
Chrome	561477	37.7	5.5
Big Rocks	8550110	5.8	1.1
Iron Ore	1020900	13.5	2.5
Cobalt	132	1.6	2.6

10. Choose **Numeric** from the **Style** menu.

    The locate region you just defined becomes a numeric region.

To make a table from the selected regions:

1. Choose **Select All** from the **Edit** menu.

   Pro OCR selects all of the locate regions you defined.

2. Choose **Make Table** from the **Edit** menu.

   Pro OCR creates a table from the selected locate regions.

3. Click the **Recognize** button in the Gallery toolbar.

   Pro OCR recognizes the page image using the locate regions you defined in the previous steps.

   After Pro OCR is finished recognizing, the page appears in the text view.

Mineral (name)	Volume (tons)	Value (US $ mil)	% Total Production (by value)
Gold	114	208.7	51.5
Asbestos	186581	51.0	9.9
Nickel	11489	108.9	22.1
Lint	16166	34.0	6.6
Chrome	561477	37.7	5.5
Big Rocks	8550110	5.8	1.1
Iron Ore	1020900	13.5	2.5
Cobalt	132	16	2.6

You have completed this example. A message appears asking if you want to save the document.

4. Choose **Close** from the **File** menu.

   Close the document without saving it.

© Copyright 1998 Visioneer, Inc. Reach us at www.visioneer.com.

Chapter 3

Getting Documents

This chapter tells you how to get (acquire) documents with Pro OCR. It is assumed that you completed the procedures in "Starting Pro OCR," and "Selecting a TWAIN-Compliant Scanner," in Chapter 2.

In this chapter you learn:

- The basic steps for getting a page
- How to get a page using a scanner
- How to get a page from a file

Getting a Page—The Basic Steps

There are two ways to get a page: 1) Use Auto OCR to automatically get a page. 2) Perform an individual Get Page. In each case you need to select the source—your scanner or an image file—that you want to use to get the page. If you select a scanner, you also need to select a few other options. The following procedure tells you the basic steps to get a page. For more detailed information, see "Getting Pages From a Scanner," and "Getting Pages from an Image File," later in this chapter.

To get a page:

1. Select a source from the **Get Page** drop-down list, in the Gallery toolbar.
2. If you select Use Scanner, select scanner options as described in"Setting Scanning Options," later in this chapter.
3. Click **Get Page** or **Auto OCR** depending on which process you want to use.
4. If you select Use Scanner, scan the document using your scanner. If you select Open File, open the file you want to use.

Getting Pages From a Scanner

You can use a scanner to get one page at time by using the Get Page button, or use a scanner with Auto OCR to get multiple pages automatically. This section tells you how to:

- Set scanning options
- Get one page using Get Page
- Get pages with Auto OCR

Setting Scanning Options

To set scanner settings for your scanner, such as the resolution, brightness, and page orientation, see the documentation that came with your scanner. You can set the following processing options in Pro OCR by choosing Options from the Tools menu:

- **Straightening Skewed Images.** Automatically straightens type that is skewed (crooked) on a page. When text on a page is badly skewed, Pro OCR may have trouble correctly locating paragraph boundaries. Recognition may also be affected, resulting in many illegible characters.

  NOTE: Processing with the Straighten Skewed Images option selected takes longer than processing the same page with this option not selected. However, recognition is usually much better on skewed type if the page image has been straightened. You may want to experiment on skewed pages to see when to use the Straighten Skewed Images option. Pro OCR is preset to not straighten skewed images.

- **Splitting one A3 page.** For scanners that scan two, 11 by 17 inch pages, you can scan bound material and Pro OCR will automatically split the image into two pages.

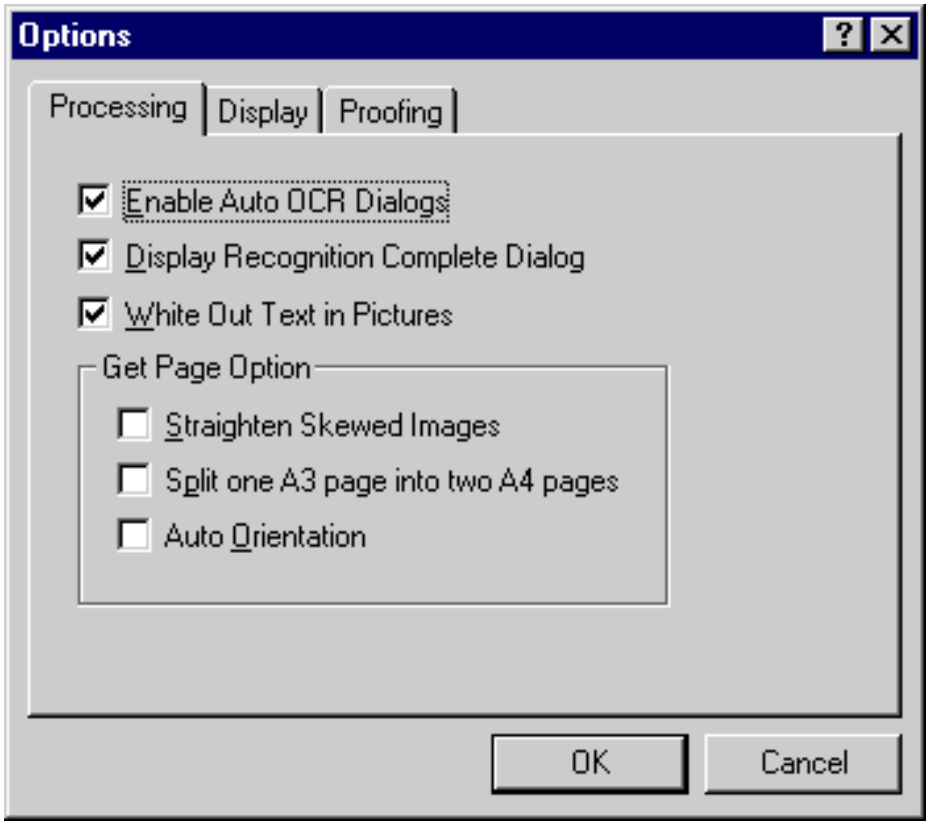
- **Auto Orientation.** Automatically selects Portrait or Landscape orientation for the page.

By default, Pro OCR does not select these settings for you.

To set Get Page Processing options:

1. Choose **Options** from the **Tools** menu.

   The Options dialog box appears with the Processing tab selected.



2. Select the options that you want to use.

3. Click **OK**.

Selecting a Scanner as the Source

When you get pages from a scanner by using Auto OCR, a deferred job, or Get Page, one or more page images are read in from the scanner. Pages are scanned according to the current page size, orientation, brightness, and scanning settings selected in your scanner's software.

When you read in additional pages from a scanner, new page images are added to the active document. You can read up to 999 pages into a document, as long as you have enough available disk space.

To select a scanner as the source:

- Select **Use Scanner** from the **Get Page** drop-down list in the Gallery.

NOTE: If you did not previously select a scanner, the Select Scanner dialog box appears, letting you select one now. (You can also select a scanner by choosing Select Scanner from the Tools menu.)

Getting a Page Using a Scanner

During the single-step Get Page operation, you scan only one side of one page at a time. You cannot automatically read stacks of pages or double-sided pages. Instead, you must manually feed pages that you want to be read. The procedure is the same whether you use an automatic document feeder or a flatbed scanner.

When you scan in a page using Get Page, the new page is added after the current page. If you want to add pages to the end of a document, make sure the last page of the document is displayed before you do Get Page. To insert a page after any other page, make sure the appropriate page is displayed. Go to the page, if necessary, and then use Get Page to insert the new page after it. You can also use single-step Get Page to replace a current page.

To get one page from a scanner using Get Page:

1. Make sure you have set scan options as described in "Setting Scanning Options," previously in this document, and select **Use Scanner** from the **Get Page** drop-down list.

2. If you are adding pages, to other pages that you already got, make sure the current page is displayed in Pro OCR.

   New pages are added after the current page.

3. Place one page on the flatbed or place one page in the ADF.

   Make sure the page is oriented correctly for your scanner and the orientation you have selected.

   You can put in as many pages as the ADF will hold, but Pro OCR will only scan one page at a time using Get Page.

4. Click **Get Page**.

   The Get Page button is highlighted to indicate that Pro OCR is getting pages. In the status display area, a meter bar indicates that Pro OCR is scanning the page.

Pro OCR scans the page on the flatbed or the first page in the ADF, using the current brightness, page size, orientation, and scanning resolution settings. After the single page is read in, it appears using the previous magnification.

NOTE: To find the most appropriate brightness setting for a page, use Get Page to scan the same page as many times as necessary. You can change the level of brightness in your scanner's software.

To scan additional pages:

1. Place another page on the flatbed.

2. Click **Get Page**.

   Repeat steps 1 and 2 for each additional page you want to scan.

NOTE: After Get Page is completed, whether or not pages have been located or recognized, you can save files in Pro OCR format, or in any of the other image output file formats. for more information about saving, see Chapter 6, "Saving and Printing Documents."

Using Auto OCR with Scanners

This section tells you how to use Auto OCR with a flatbed scanner or Automatic Document Feed (ADF) scanner.

NOTE: When scanning pages, make sure that pages are placed as straight as possible. Pages skewed more than 2° may result in the incorrect sorting and grouping of text lines unless the Straighten Skewed Images processing option is selected. Also note that pages skewed more than 0.5° may jam in an ADF.

Using Auto OCR with a Flatbed Scanner

To use Auto OCR with a flatbed scanner, complete the following procedure.

NOTE: You cannot scan double-sided pages automatically when using a flatbed scanner. You should place the pages on the scanner's bed in the order in which you want the text to be read.

To automatically process one or more pages with a flatbed scanner:

1. Make sure you have set scan options as described in "Setting Scanning Options," previously in this document, and select **Use Scanner** from the Get

Page drop-down list.

2. Check the Locate and Recognize options to make sure they are set the way you want them.

3. Place the first page on the flatbed.

   Make sure the page is oriented correctly for your scanner and the page orientation you have selected in the Gallery.

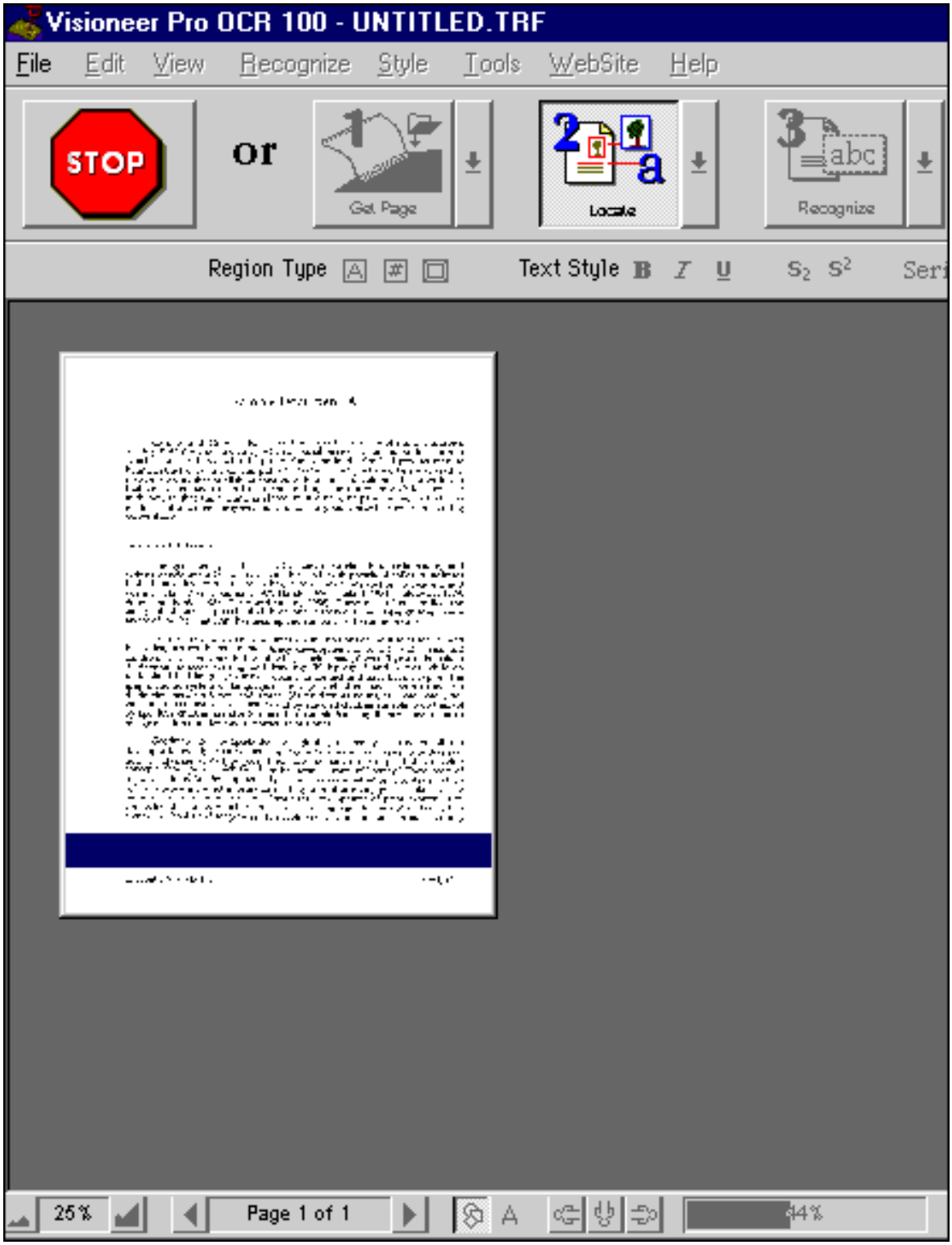
4. To scan more than one page, choose **Options** from the **Tools** menu, and then select the **Enable Auto OCR Dialogs** processing option.

5. Click **Auto OCR**.

   The scanner software appears.

6. Use the software as you usually do.

   Pro OCR begins getting pages:



If the Enable Auto OCR Dialogs processing option is not selected, scanning is completed. Pro OCR begins locating and then recognizing.

If the Enable Auto OCR Dialogs processing option is selected, Pro OCR asks for additional pages to scan after it finishes reading in the current page:

7. If you want to get additional pages, place another page on the flatbed.

   Pro OCR scans the additional page on the flatbed and displays the dialog box again, asking for the next page. Repeat this step for as many additional pages that you want to scan.

8. If you do not want to scan more pages, click **End**.

   Scanning is completed. Pro OCR displays the page you've scanned in the image view. Pro OCR then begins locating and then recognizing.

Using Auto OCR With a Scanner with an ADF

Complete the following procedure to use Auto OCR with scanners that have an ADF.

NOTE: To use an ADF scanner with Pro OCR, you need the Pro OCR ISIS upgrade. For more information, visit Visioneer's Web site at www.Visioneer.com.

To automatically process one or more pages with a scanner that has an ADF:

1. Make sure you have set scan options as described in "Setting Scanning Options," previously in this document, and select **Use Scanner** from the Get Page drop-down list.

2. Place one or more pages in the ADF.

   Make sure the pages are oriented correctly for your scanner and the page orientation you have selected in the Gallery.

3. To scan more than one page, choose **Options** from the **Tools** menu, click the **Processing** tab, and then select the **Enable Auto OCR Dialogs** processing option.

4. Check the Locate and Recognize options to make sure they are set the way you want them.

5. Click **Auto OCR**.

   Pro OCR begins getting pages.

   If the Enable Auto OCR Dialogs processing option is not selected, scanning is completed. Pro OCR begins locating and then recognizing.

   If the Enable Auto OCR Dialogs processing option is selected, Pro OCR asks for additional pages to scan.

6. If you want to scan another stack of pages, place the next stack of pages in the ADF.

   Pro OCR scans the additional pages in the ADF and displays the dialog box again, asking for additional pages to scan. If you need to scan the second side of a stack of double-sided pages, see the next procedure, "To scan the second side of double-sided pages:."

   Repeat this step for as many additional stacks of pages as you want to scan.

7. If you've scanned all the pages you need for this job, click **End**.

   Scanning is completed. Pro OCR displays the first page of the scanned stack, in the image view. Pro OCR then begins locating and recognizing.

To scan the second side of double-sided pages:

1. When you're finished scanning the first side, turn the entire stack of pages in the ADF over and replace them in the ADF.

   Make sure that you don't change the order of pages, and that you replace them in the proper orientation. If your double-sided document contains more pages than your ADF can handle, you'll need to separate the document into smaller stacks. After scanning the first side of a smaller stack, scan the flip side of the same stack before continuing with the next stack.

2. Click **Flip** in the dialog box that appears.

   Pro OCR scans the second side of each page using the current brightness, page size, orientation, and scanning resolution settings.

3. When you're finished, click **End**.

Scanning is completed. Pro OCR finishes getting pages and displays the first page of the scanned stack in the image view. The scanned double-sided text is correctly sequenced, in correct page order.

Getting Pages from an Image File

Typically, Pro OCR obtains the image of a page by working directly with your scanner. You can, however, also use Pro OCR with image files you scanned or created using other applications. There are several common sources for obtaining image files, other than scanning with Pro OCR:

- Scanner applications not supported by Pro OCR
- Fax-modem applications
- High-resolution paint programs

Pro OCR can read the following image file formats:

- TIFF (Uncompressed, PackBits, Group 3, Group 3 modified, Group 4)
- PCX
- DCX

Pro OCR can open black-and-white (one-bit) single-page or multiple-page image files. Pro OCR does not open grayscale (greater than one-bit) or color image files.

Not all instances of the above files from every application are supported, however, because specific implementations of these formats are not necessarily standard. If you try to open a file of a type that Pro OCR doesn't recognize, Pro OCR displays a warning message.

Selecting a File as the Source and Getting Pages

The following procedure tells you how to select and open a file as the source for Get Page.

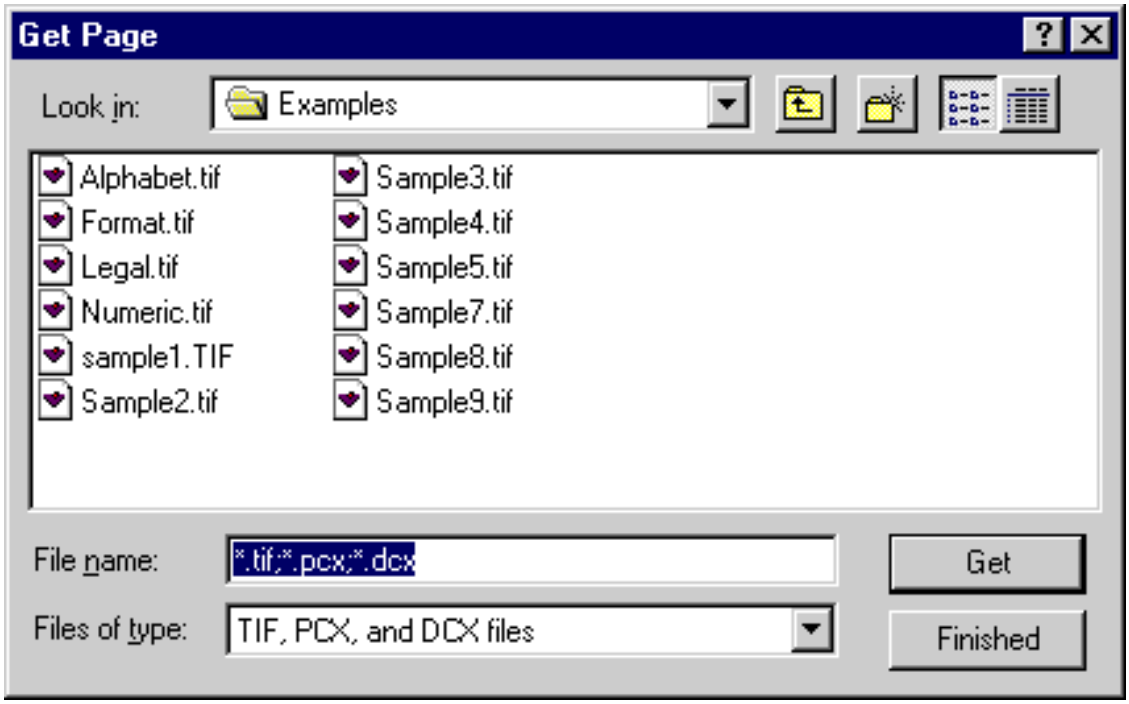
To select and open files:

1. Select **Open File** from the **Get Page** drop-down list.

   A checkmark appears next to it when selected.

2. Click the **Get Page** button in the Gallery toolbar.

   The Get Page dialog box appears.



3. Select the file and click **Get**.

   The file is read in and the progress bar moves down the page.

Getting Files From Other Scanner Applications

Pro OCR supports many of the most popular scanners directly. However, if you don't have a scanner that Pro OCR supports directly, you may still be able to use Pro OCR with the scanner application you do have. Most scanner applications save to one of the image file formats that Pro OCR supports.

To get pages from a non-supported scanner:

1. Scan a page using a scanner application that is compatible with your scanner.

2. Save the page in an image file format that Pro OCR supports.

3. Select **Open File** from the **Get Page** drop-down list.

4. Click **Auto OCR**.

5. Find and select the file(s) that you want to process.

6. Click **Add** and then click **Get**.

   Pro OCR automatically processes the image file(s) according to the controls in the Locate and Recognize rows of the Gallery.

   OR

1. Click **Get Page**.

2. Find and select the file that you want to process.

3. Click **Add** then click **Get**.

   Pro OCR reads in the specified file. You can continue with any combination of the single-step Locate and Recognize operations, followed by Finish Processing, or save it in the Pro OCR Deferred format and finish processing it later on using Process Deferred Jobs.

   After the page is read in, Pro OCR treats the page as if it had scanned it.

Getting Fax-modem Files

Pro OCR can also open fax-modem files, if they have been saved in one of the supported input file formats.

Many fax-modems have both a Standard and a High-Resolution (or Fine) setting. The Standard setting typically transmits characters at 204 x 98 dpi. The High-Resolution setting typically transmits at 204 x 196 dpi. Fax-modem files transmitted at Standard setting may not be recognized by Pro OCR as accurately as those transmitted at High-Resolution.

To get a fax-modem file, use the same procedure as described in the previous section.

NOTE: It is recommended that you use the highest resolution a fax-modem can produce for the best possible recognition.

Using Auto OCR With a File

You can specify one or more image files for the Get Page step, and then have Pro OCR automatically locate and recognize them. If you've selected the Enable Auto OCR Dialogs processing option, you can also select one or more additional files after reading the initial files and before locating and recognizing begin. Pro OCR can process most standard black and white TIFF, PCX, and DCX files.

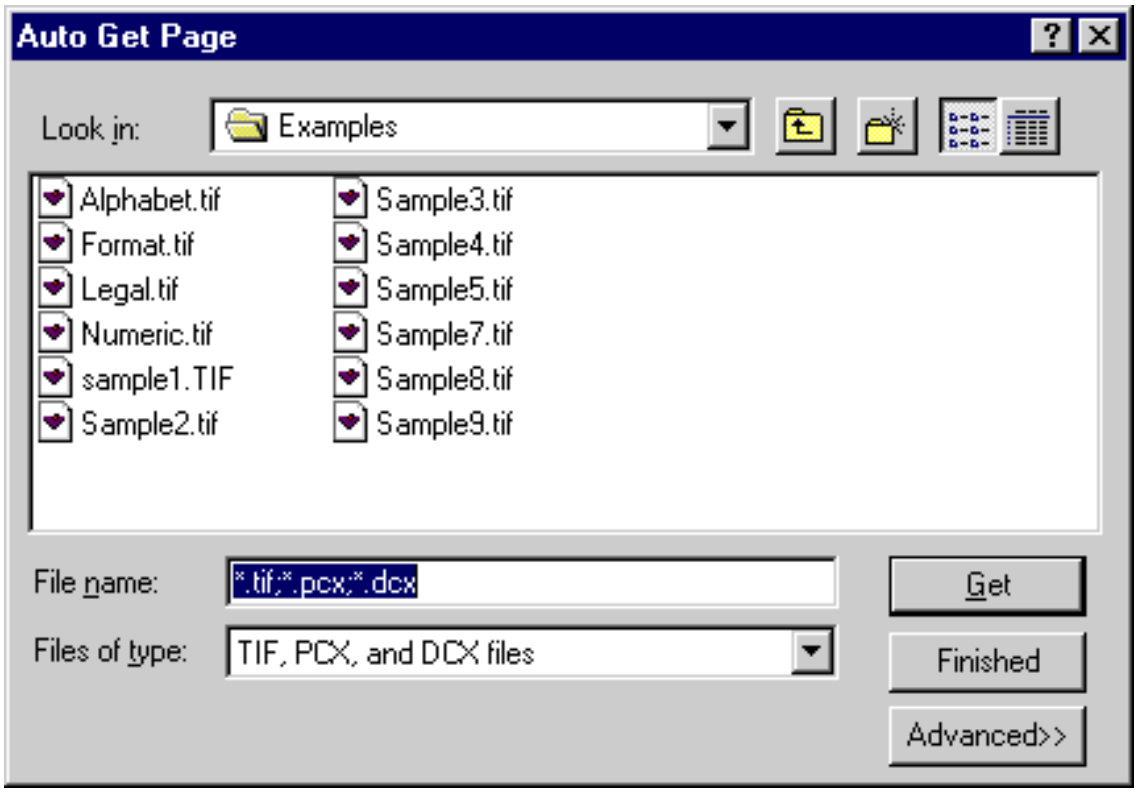
To automatically process from a file:

1. Select **Open File** from the **Get Page** drop-down list in the Gallery toolbar.

2. Check the Locate and Recognize options to make sure they are set the way you want them.

3. Click **Auto OCR**.

   The following dialog box appears:



4. Find and select the files you want.

   To get just one file, click the file name.

   To get multiple files, click the Advanced button. The dialog box expands. Click the file that you want to get and then click the **Add** button. The file names appear in the Selected Files list in the lower half of the dialog box. You

can add available files from as many directories and disks as necessary. Files are displayed in the Selected Files list in the order in which you add them.

NOTE: To remove a file from the Selected list, select the file name and click the **Remove** button. To remove all selected files, click Remove All.

5. Click **Get**.

   Pro OCR reads in the selected file or files. As a page is read in, the Get Page button is highlighted, and the progress bar moves down the page.

   Each page is read in and displayed in the image view at 25% magnification (zoom level).

   If the Enable Auto OCR Dialogs processing option is not selected, when all pages have been read Pro OCR finishes getting pages and displays the first page in the image view. Pro OCR then locates and recognizes each page.

   If the Enable Auto OCR Dialogs processing option is selected, when all pages have been read the Get Page dialog box is again displayed.

6. To add pages from an additional file or files to the end of your current document, repeat steps 5 and 6 as often as necessary.

   Each time you read in another file, the new pages are read in and added to the end of the current document. You can read up to 999 pages into a document, as long as you have enough available disk space.

7. When you're done reading files, click **Finished**.

   When you click End, the file reading step completes, and locating and then recognizing begins. For more information about locating see Chapter 4, "Locating Text and Graphics." For more information about recognizing and proofing, see Chapter 5, "Setting Recognize Options and Proofing a Recognized Document."

NOTE: When you use Auto OCR, the locate and recognize steps occur automatically.

More About Enabling Auto OCR Dialogs

By default, after you've used Auto OCR to scan pages or to read in one or more files, Pro OCR displays a dialog box that prompts you to continue in one of several ways:

- Scan another page or stack of pages
- Scan the second side of a page or stack
- Open additional files

This lets you read in and process multiple files or stacks of pages as one document.

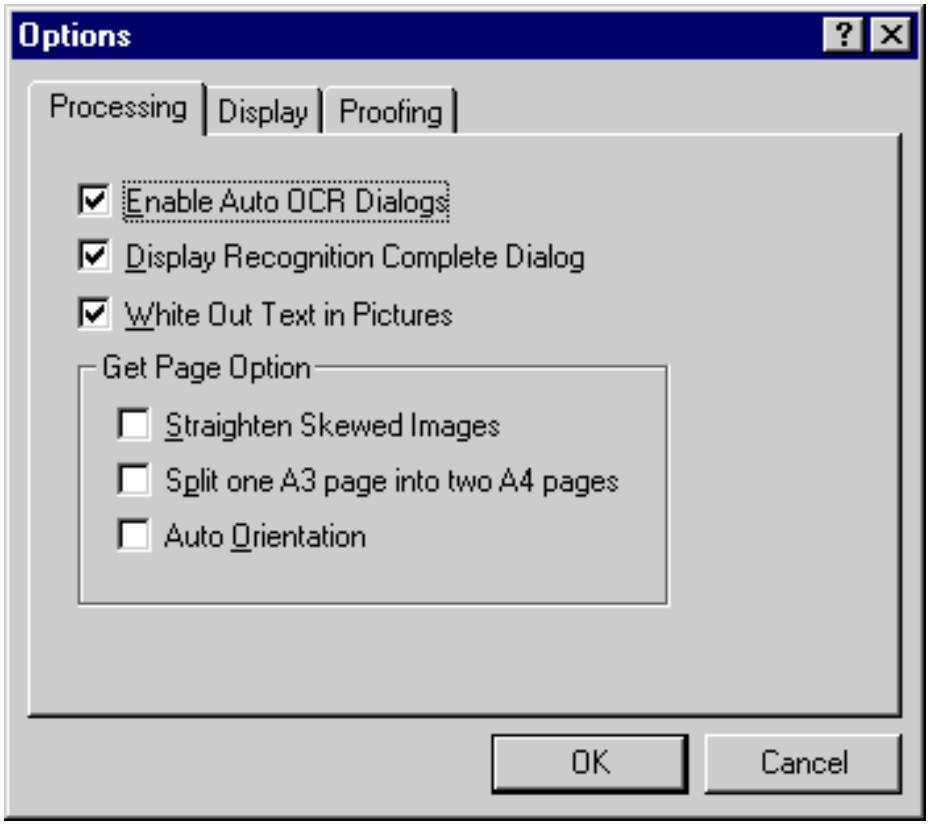
However, it also means that you have to click Finish to proceed with automatic processing after the Get Page step is done. If instead you want the Auto OCR process to continue without interruption, you can prevent the dialog box from reappearing by deselecting the Enable Auto OCR Dialogs option.

NOTE: You can't process more than a single stack of pages or set of files when Enable Auto OCR Dialogs is deselected.

To enable/disable Auto OCR dialog boxes:

1. From the **Tools** menu, choose **Options**.

   The Options dialog box appears with the Processing options.



2. To enable the dialogs, select **Enable Auto OCR Dialogs**. To disable the dialog boxes, deselect the option.

© Copyright 1998 Visioneer, Inc. Reach us at www.visioneer.com.

Chapter 6

Saving and Printing Documents

This chapter describes the input file formats and output file formats that Pro OCR supports and tells you how to save documents in a variety of these formats.

Saving Documents and Other Pro OCR Items

You can save the following documents and items:

- Documents (in various file formats)
- Templates (text, numeric, picture, and table region definitions and ordering information)
- Gallery settings and selected processing, display, and proofing options

Saving a Document

Documents are not saved automatically. You save a document using Save or Save As from the File menu. If you close or exit Pro OCR without saving a document, a message prompts you to save the current document.

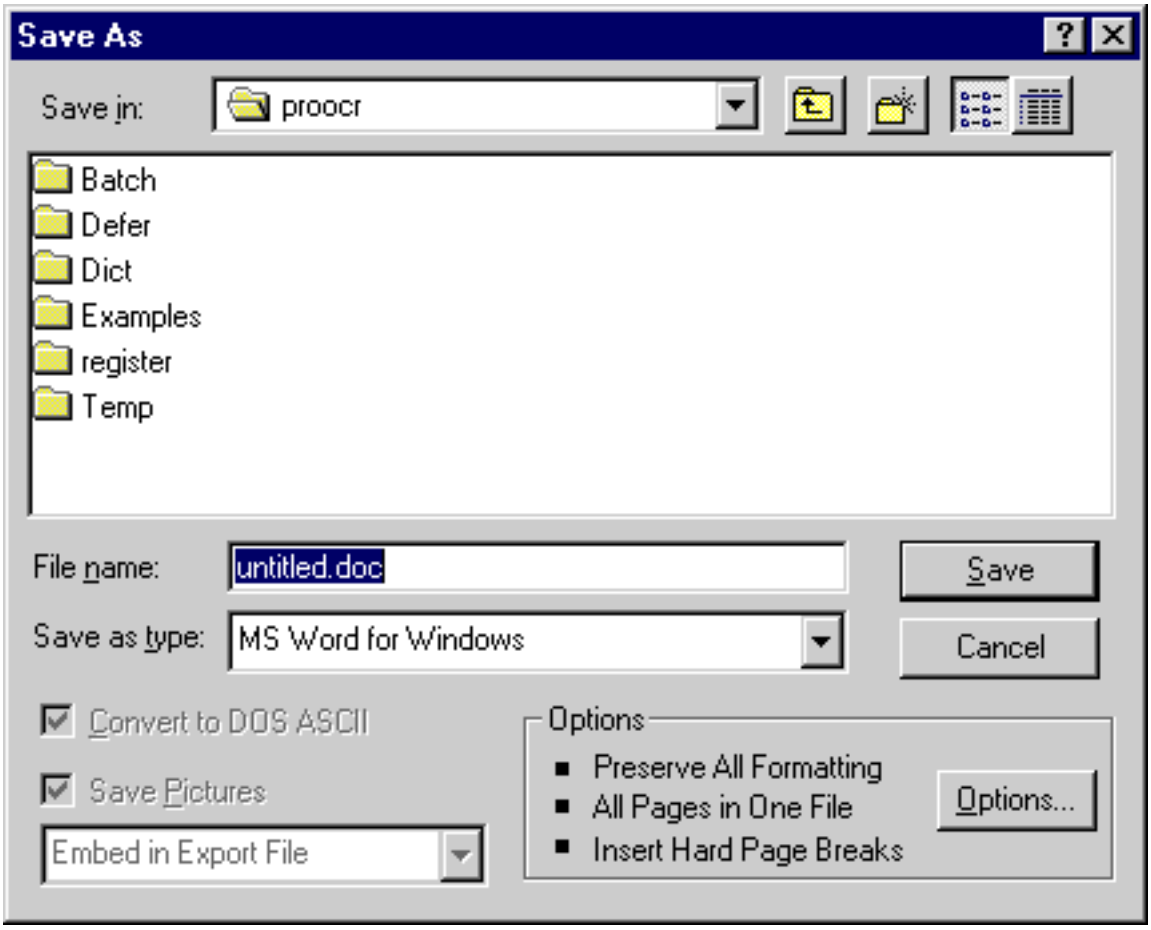
After you get a document, you can save it at any or all of the various stages of the Pro OCR process—after locating, recognizing, or proofing. If a document does not contain recognized text, you can save only as Pro OCR, Pro OCR Deferred, or using one of the standard image file formats.

NOTE: When you save to formats other than Pro OCR or Pro OCR Deferred, you must still save the document in one of the Pro OCR formats to be able to use it again in Pro OCR.

To save an open document:

1. Choose **Save As** from the **File** menu.

The Save As dialog box appears:



If the document has been saved previously, the name of the document is displayed and selected in the File Name box. If the document has not already been saved, the File Name box is selected and contains the default file name: `UNTITLED.XXX`. Pro OCR adjusts the file extension represented here as `XXX` according to the document format you select in the Save as Type drop-down list.

2. Type a new file name, if necessary.

3. Choose a document format from the **Save as Type** drop-down list.

   If this is a new file, the last used document format is displayed. If this is a previously saved file, the previously saved document format is displayed.

   You can choose from the following document formats:

   - Pro OCR document file formats
   - Standard image file formats

- Standard text file formats
- Word processor and spreadsheet file formats

For more information about the different file formats, see “Supported Output File Formats” later in this chapter.

4. If you want to save any pictures in the document, select the **Save Pictures** option and choose a picture format from the **Picture Format** drop-down list.

   NOTE: Saving pictures in a document is different from saving the entire page image. Save the page image using one of the image file formats presented in the Save as Type drop-down list. For more information, see “Saving Pictures” later in this chapter.

5. If you want to embed any pictures into the document when it is saved, choose **Embed in Export File** from the **Picture Format** drop-down list.

   The embedding option is only available for the following document formats:

   - MS Word for Windows
   - Rich Text Format (RTF)
   - WordPerfect 5.0 and 5.1

   OR

   If you want to save the page images only, choose one of the other picture formats from the Save as Type drop-down list.

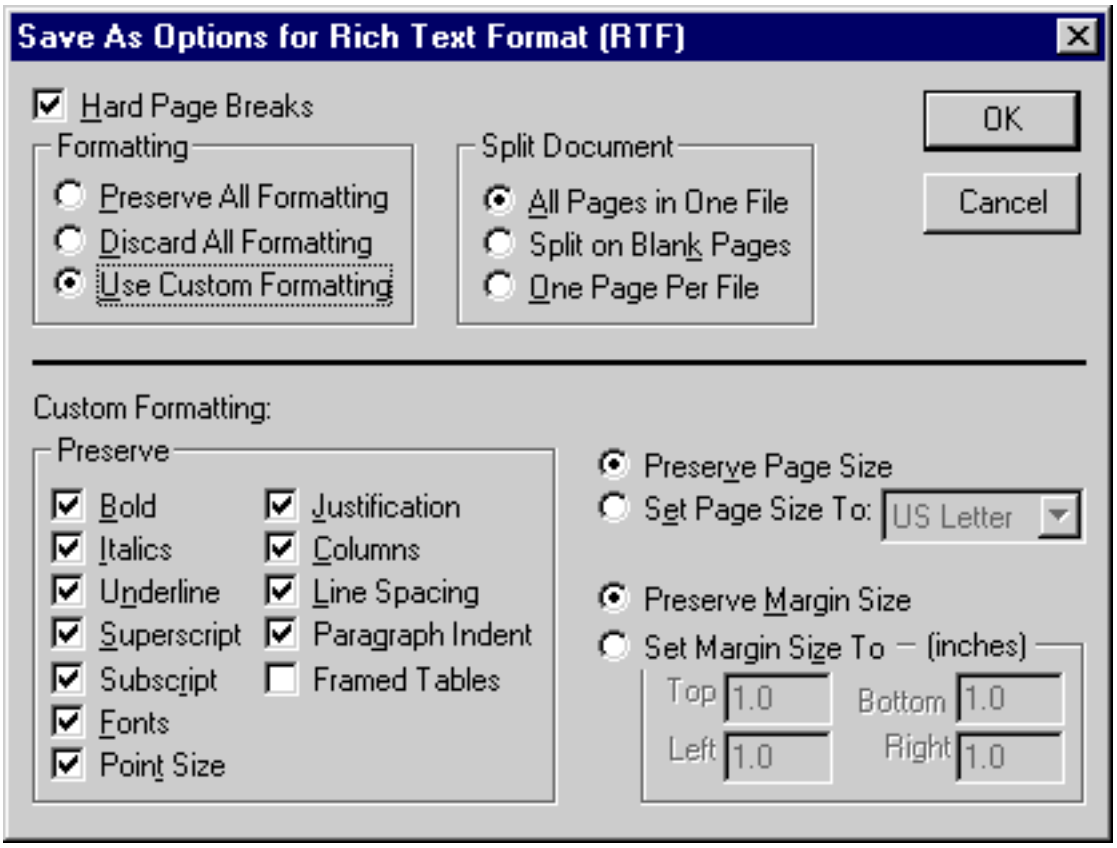
   Choosing a picture format tells Pro OCR to save only the page images from the active document.

   NOTE: When saving the document in one of the standard TIFF formats, you can choose whether to save all pages in one file, to split on blank pages or to save one page per file. When saving to the PCX format, you must save one page per file.

6. To select the formatting information that will be exported to the format

currently chosen in the Save as Type drop-down list, click the **Options** button to open the Save As Options dialog box.

Most formats have additional options. If there are no options available for the format you've selected, the Options button is dimmed.



The Save As Options dialog box has the following sets of options:

- If page breaks should be inserted between each page
- If formatting should be preserved or completely discarded, or if only certain formatting should be preserved
- If all pages in the document should be saved as a single file, or as separate files for each page

If you decide to only save certain formatting, you can select from the following formatting to be saved:

- Style
- Font (typeface)

- Point size
- Justification
- Number of columns
- Line spacing
- Paragraph indentation
- Page size
- Margin sizes

Choose one of the Split Document options to either keep all pages in one file or split the document into multiple files:

- **All Pages in One File:** Choose this option to save all the pages in the document in one file.
- **Split on Blank Pages:** Choose this option when you want Pro OCR to save a stack of documents into separate files.

To use this option, before you scan the stack of pages, put a blank page after the last page of each document you want Pro OCR to save as a separate file. For more information, about saving multiple documents, see “Saving Multiple Documents as Separate Files” and see “Saving Multiple Page Images as Separate Image Files” later in this chapter.

NOTE: For Split on Blank Pages to work properly, make sure to use the Recognize operation on every blank page.

Pro OCR saves each stack of pages up to a blank page as a separate file, using the name you specified followed by a sequential three- digit numeric identifier, followed by the appropriate extension. For example, if you name the current document `BOOK`, and then save it to Excel 2.x format with “Split on Blank Pages” selected, Pro OCR will save the first file (up to the first blank page) as `BOOK001.XLS`, the next file as `BOOK002.XLS`, and so on.

- **One Page Per File:** Many image editing programs can support only

one image page per file. If you save in PCX format, Pro OCR automatically selects this option, because a PCX file can only have one page.

When you use this option, Pro OCR automatically creates one file for each page. Pro OCR saves each file using the name you specified followed by a sequential three-digit numeric identifier, followed by the appropriate extension. For example, if you name the current document `IMAGE`, and then save it in a TIFF format with "One Page Per File" selected, Pro OCR saves the first page image as `IMAGE001.TIF,` the next page image as `IMAGE002.TIF`, and so on.

7. If you opened the Save As Options dialog box, click **OK** to close it.

   The Save As dialog box reappears.

8. Click **OK**.

   The document is saved according to the selected options.

   If you try to save the document with a name that has already been used, a dialog box asks if you want to replace the existing document. Click No to return to working with the document. Click Yes to replace the document.

NOTE: When you want to open a document in an image editing program, save it in one of the image file formats. Any locate regions that have been applied or created are not saved. If the document has been recognized, the recognized text is not saved.

Saving Multiple Documents as Separate Files

Often you'll have many documents on which you want to do Get Page, Locate, and Recognize at one time, but you want the recognized files saved as separate documents. Pro OCR makes it easy for you to process a large stack of separate documents as one and still keep them separate when you save them. You can do this when you're saving to a text format, the various image output formats, or to any export format.

To save multiple multipage documents as separate files using the split option:

1. Before you put the pages in the scanner, separate the documents by putting a blank piece of paper between each document and the next.

2. Process the pages as you would normally.

3. When you save the document, choose **Save As** from the **File** menu.

   The Save As dialog box appears.

4. Click the **Options** button.

   The Options dialog box appears.

5. Select the **Split on Blank Pages** option and click **OK**.

   Pro OCR saves each stack of pages up to a blank page as a separate file, using the name you specified followed by a sequential numeric identifier, followed by the appropriate extension. For example, if you name the current document `BOOK`, and then save it to Excel 2.x format with the Split on Blank Pages option selected, Pro OCR will save the first file (up to the first blank page) as `BOOK001.XLS`, the next file as `BOOK002.XLS`, and so on.

6. Click **OK**.

To save multiple single-page documents as separate files using the one page option:

1. Process the pages as you usually would.

2. When you save the document, choose **Save As** from the **File** menu.

   The Save As dialog box appears.

3. Click the **Options** button.

   The Options dialog box appears.

4. Select the **One Page Per File** option and click **OK**.

   Pro OCR saves each page image as a separate file, using the name you specified followed by a sequential numeric identifier, followed by the appropriate extension. For example, if you name the current document `PAGES`, and then save it to RTF format with the “One Page Per File” option selected, Pro OCR will save the first page image as `PAGES001.RTF`, the next page image as `PAGES002.RTF`, and so on.

5. Click **OK**.

Saving Multiple Page Images as Separate Image Files

In addition to Pro OCR format, you can save a document in a number of image output formats. Usually, you'll save a copy of your document in one of these graphic formats when the document you're processing has illustrations that you want to save and use in other applications. Because many image-processing programs cannot process multipage image files, you'll probably want to save multipage documents one image per file.

To save multiple pages as separate image files:

1. Process the pages as you usually would.

2. When you save the document, choose **Save As** from the **File** menu.

   The Save As dialog box appears.

3. Click the **Options** button.

   The Options dialog box appears.

4. Select the **One Page Per File** and click **OK**.

   When you name the file, choose a file name of up to five characters. If the file name is longer, Pro OCR truncates it to five characters.

   Pro OCR saves each page image as a separate file, using the name you specified followed by a sequential numeric identifier, followed by the appropriate extension. For example, if you name the current document `IMAGE`, and then save it to PCX format with "One Page Per File" selected, Pro OCR will save the first page image as `IMAGE001.PCX`, the next page image as `IMAGE002.PCX`, and so on.

Saving Templates

Save a template when you've defined locate regions that can be applied to other page images. A template may be used to identify the locate regions on all pages to be recognized. Or, you can use different templates to identify locate regions on different pages.

To save a template:

1. Choose **Save Template As** from the **File** menu.
2. Enter a file name.
3. Choose the format from the **Save as Type** drop-down list, and then click **Save**.

   You can open a saved template by double-clicking the Template button or by choosing **Select Template** from the **File** menu.

Saving Settings

The current settings are remembered when you open Pro OCR again. You can also save settings using Save Settings As from the File menu, and open those settings later when you need them.

To save settings:

1. Choose **Save Settings As** from the **File** menu.
2. Enter a file name.
3. Select a format type from the Save as Type drop-down list.
4. Click **Save**.

To retrieve settings:

1. Choose **Retrieve Settings** from the **File** menu.

   The Save Settings As dialog box appears.
2. Select the settings file you want to use.
3. Click **Open**.

Supported Output File Formats

This section provides additional details about the different Pro OCR output file

formats. Pro OCR can save to a variety of output file formats at various stages of processing.

Table 6-1: Proprietary Pro OCR Formats

Pro OCR	Pro OCR Text Only
Pro OCR Deferred	

Table 6-2: Standard Image File Output Formats

TIFF Uncompressed	TIFF Group 3
TIFF PackBits	TIFF Group 3 Modified
PCX	TIFF Group 4

Table 6-3: Standard Text File Formats

Plain Text	Formatted Text
Text with Line Breaks	Comma Delimited Text
Tab Delimited Text	Rich Text Format (RTF)
HyperText Markup Language	

Table 6-4: Word Processor File Formats

Lotus Ami Pro	WordPerfect 5.x
Microsoft Word for Windows	

Table 6-5: Spreadsheet File Formats

Microsoft Excel	Lotus 1-2-3

NOTE: If you don't have any of the applications listed here, note that many word processor and spreadsheet applications can handle formats from other word processors and spreadsheets. Most Windows word processors can import RTF files, although some have only limited support for RTF.

Internally, Pro OCR preserves all the format, character style, and font information of the input page. What is actually retained when you export the file depends on four things:

- The formatting options you choose
- The document format you save to
- The picture formats you save to
- The application you open the saved file in

As long as there are pages in your document, you can save it in various TIFF and PCX formats, in the Pro OCR format, and in the Pro OCR Deferred format. As soon as any pages have been recognized, you can save the document in all supported file formats.

NOTE: If you save in the Pro OCR Text Only format, you won't be able to use the On-Screen Verifier during editing when the file is opened again.

Saving to Proprietary Pro OCR Formats

You save to the Pro OCR, Pro OCR Text Only, or Pro OCR Deferred file format when you want to open and process a document again in Pro OCR. When you save to any text or application formats, you must still save the document in one of the Pro OCR formats to be able to open it again in Pro OCR.

Saving to Pro OCR Format

This file format retains all document information needed to subsequently create any of the other supported file formats.

There are several reasons why you might want to save a copy of your document in Pro OCR format. The primary reason is to save work in progress so that you can open the document later for further processing.

The current state of each page in the document is saved, including any locate regions or recognized text.

Saving to Pro OCR Deferred Format

The Pro OCR Deferred file format is a special case of the Pro OCR file format. Use it to save work in progress so that you can open the document later for further single-step processing (using the Open command in the File menu) or to complete processing (using the Process Deferred Jobs command in the File menu). A document is automatically saved in the Pro OCR Deferred file format when you choose Create Deferred Job from the File menu. You can also save the open document as a deferred job using the Save As dialog.

Using Open and Process Deferred Jobs with Pro OCR Formats

There are two things you can do with a file saved in Pro OCR format or in Pro OCR Deferred format:

- Read it in for further single-step processing using Open.
- Complete automatic processing on it using Process Deferred Jobs.

When you use Open to read in a file that you've saved in the Pro OCR format or the Pro OCR Deferred format, each page is retrieved with the Locate and Recognize information that was saved with it.

In contrast, if you use the single-step Get Page operation to read in a file saved in Pro OCR format or in Pro OCR Deferred format, only the page image is read in, and the locate regions and recognized text in the document are discarded.

NOTE: When you're reading a file in for further processing, remember to use Open instead of Get Page in order to preserve any locate regions or recognized text you've already processed.

Saving to Pro OCR Text Only Format

The Pro OCR Text Only format preserves the text as you see it displayed in the text view in Pro OCR. It also preserves all current information about suspect and illegible characters in the recognized document. It lets you open the document again in Pro OCR without the page image.

When you open the document later on, you can edit the text, continue to inspect

suspect and illegible characters if there are any left, check spelling, and search for numbers, punctuation, symbols, and alphanumeric words. However, because you haven't saved the page image, you can't use the On-Screen Verifier.

Text files take up a lot less space than image files. Image files are large, even when compressed.

Saving to Standard Image File Formats

You can save a document in a variety of TIFF formats and in PCX format. Only page images are saved, even if pages have already been located or recognized. You can save to these formats at any time, as long as you have pages in your document.

NOTE: When you use Save As to save to formats other than Pro OCR, Pro OCR Text Only, or Pro OCR Deferred format, you must still save the document in one of the Pro OCR formats to be able to use the text again in Pro OCR.

You can open a document saved in TIFF or PCX format in many image editing applications.

NOTE: Many image editing programs can only support one image page per file, For this reason, the Save As command has an option that lets you save a multipage document as a sequence of single-page TIFF files. PCX files are always saved as one image page per file.

Saving to Generic Text File Formats

You can save in the generic text formats only after recognizing. The following formats are general purpose text formats that many word processor, spreadsheet, and database applications can import, either directly or using a filter or conversion process.

NOTE: Windows and Windows applications use the ANSI standard for representing text in text files. DOS and DOS applications use the ASCII standard. If you save in one of the text formats (Plain Text, Text with Line Breaks, Formatted text, Tab Delimited Text, or Comma Delimited Text) and plan to use this text in a DOS application, make sure to select the Convert to DOS ASCII option in the Save As dialog box.

- **Plain Text.** Preserves text, tabs, and carriage returns at the ends of paragraphs. No page formatting, character style, or font information is

preserved. When you output a recognized document in Plain Text format, the text is sequentially output in the order in which the text blocks were located. Margins and columns are not preserved.

- **Text with Line Breaks.** Preserves text, tabs, and a carriage return at the end of each line. No page formatting, character style, or font information is preserved. When you output a recognized document in Text with Line Breaks format, the text is sequentially output in the order in which the text blocks were located. Margins and columns are not preserved.

- **Comma-Delimited Text.** Preserves text and carriage returns, and inserts a comma wherever a tab is encountered. No page formatting, character style, or font information is preserved. When you output a recognized document in Comma-Delimited Text format, the margins and columns are not preserved.

- **Formatted Text.** Preserves text, tabs, and carriage returns. In addition, it preserves line length, margin and column information, indents, and paragraphs, using spaces where necessary. It does not preserve any other page formatting, character style, or font information.

- **Tab Delimited Text format.** Preserves text, tabs, and carriage returns. No page formatting, character style, or font information is preserved. When you output a recognized document in Tab Delimited Text format, the text is sequentially output in the order in which the text blocks were located. Margins and columns are not preserved.

- **Common Spreadsheet or Wordprocessor format.** Pro OCR provides several Save as Types for the more coming spreadsheets and word processors. See the Save As Type drop-down list in the Save dialog box for a complete selection.

- **RTF (Rich Text Format).** Preserves just about everything. A document saved in RTF format is saved with codes (or tags) that specify page format, character style, and font name and size information. When an output document is read in by an application that can decode and support the RTF codes, the output page will preserve many of the page format, character style, and font characteristics of the page you see displayed on your screen in the text view.

  NOTE: Different applications have different levels of support for RTF. Also, lines and pages may break differently in the saved document than on the screen, depending on how each word processor application deals with

font, character spacing, and line length information.

- **Hyper Text Markup Language (HTML).** Inserts HTML tags to format the document for viewing in an HTML browser.

Saving to Application Formats

When you save to a specific application format, by default Pro OCR saves as much of this format, character, and font information as possible. You can also choose to discard all formatting information, or customize the formatting that is saved with the document.

Even when you save your recognized document to a specific output format and open the document in that application, there may be differences between what you see in the text view in Pro OCR and what is displayed or printed in your application.

Format Suppression and Customizing

When you save a document with the Save As dialog box, the Save As Options dialog box lets you select a variety of format options:

- **Preserve All Formatting.** Pro OCR saves the current document with all the character formats, paragraph formats, and page formats that it was able to recognize.

- **Discard All Formatting.** Pro OCR saves the current document with no character, paragraph, or page format information. Recognized text, spaces, and tabs are preserved.

- **Custom Formatting.** You can choose which character, paragraph, and page formatting you wish to save or discard. For each attribute that you preserve, Pro OCR includes the appropriate formatting codes in the saved file. For any attributes that you choose to discard, Pro OCR does not include formatting codes, and the default formats for the word processor that you open the file in will be used.

Exporting to a Word Processor that Pro OCR

Doesn’t Support

If you have a word processor that Pro OCR does not support directly, try saving your document in one of the other Pro OCR word processor export formats. In addition, most word processors can import RTF files, although some have only limited support for RTF.

Table 6-6: Summary of Output File Formats

Format	Description	Can Pro OCR Open It Later?
Pro OCR	Saves the page image, any locate regions that have been defined, and any recognized text	Yes
Pro OCR Text Only	Saves the recognized text, including all formatting, character styling, and font information	Yes
Pro OCR Deferred	Saves the page image, any locate regions that have been defined, and any recognized text	Yes
TIFF formats	Saves the image of each page, but not the text, in the type of TIFF you select from this group	Yes
PCX	Saves the image of each page, but not the text	Yes
Standard text formats	Saves the text for each page, but not the image	No
Word processor	Saves the text for each page, and optionally, for some formats, the pictures can be embedded	No
HTML	Inserts HTML tags and saves as an HTML document for viewing with an HTML browser.	No

Saving Pictures

During the Locate and Recognize steps, if Locate Text and Pictures has been selected, Pro OCR processes any pictures, or other nontext information on the input page, as embedded graphic images. When you save to a graphic output file format or to a word processor format that supports embedded pictures, and you select the Save Pictures option in Save As, Pro OCR saves these embedded graphic images.

You can save the document without the pictures, you can save both the document and the pictures separately, and you can save the document with the pictures embedded within the document.

Printing a Document

Usually, you use Pro OCR to convert a scanned or faxed image into a text file so that the file may be opened into a word processor. After you open the recognized document in a word processor, you can print the document.

There may be times, however, when you want to print the document directly from within Pro OCR. To print a document from Pro OCR you must first complete the Recognize step. You can print from either the image view or the text view.

To print an image from Pro OCR:

1. View the document in the Image view or Text view.

2. Choose **Print** from the **File** menu.

   The standard Windows Print dialog box appears.

3. To select a different printer, change the orientation of the page, or select a specific paper size or paper tray, click on the Setup button. These selections can also be made when you go to the File menu, and select Print Setup.

4. Select options to print the entire document, the page that you are currently viewing, or a specific range of pages. You can also select how many copies of the page to print and the quality (resolution) of the print.

5. Click **OK**.

© Copyright 1998 Visioneer, Inc. Reach us at www.visioneer.com.

Chapter 4

Locating Text and Graphics

A locate region identifies an area of a page image to be recognized. You define locate regions in the image view using Pro OCR's locating procedures. This chapter tells you how to:

- Identify the different kinds of locate regions
- Select the appropriate locating method
- Locate regions automatically and manually
- Work with located regions, including redefining and deleting them

Kinds of Locate Regions

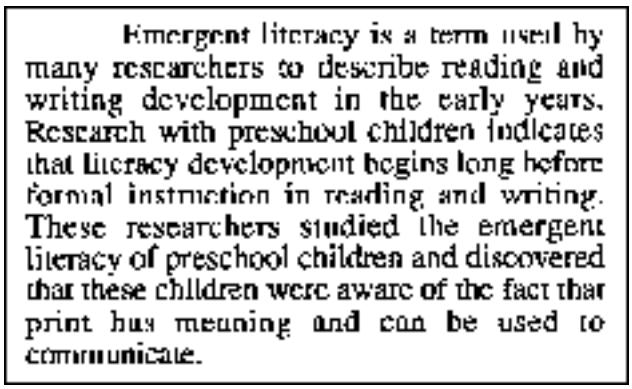
Pro OCR processes three kinds of locate regions:

- Text—contains text including letters and numbers.
- Numeric—contains only numbers and certain symbols.
- Picture—contains a picture.

Text Regions

A text region is a locate region that Pro OCR recognizes as text, including letters, numbers, and symbols. You can define text regions automatically, manually, or with a template. A selected text region can also be redefined as any other kind of locate region using the Style menu or the Style ribbon.

A single box encloses a text region:


Emergent literacy is a term used by many researchers to describe reading and writing development in the early years. Research with preschool children indicates that literacy development begins long before formal instruction in reading and writing. These researchers studied the emergent literacy of preschool children and discovered that these children were aware of the fact that print has meaning and can be used to communicate.

Numeric Regions

A numeric region is a locate region that Pro OCR recognizes as numbers (0–9) or one of the symbols shown in the following table.

Table 4-1: Numeric Symbols

+	-	¥	/	=	*	“
%	$	#	£	¢	¥	,
E	e	(	)	{	}	.
[	]	<	>	•	°	

A numeric region is enclosed in a double box with dots between the lines:

1989	1990
160	190
82	90
59	100
71	90
322	340
15	17
12	12

If Pro OCR encounters a letter of the alphabet, or a symbol other than one of the numeric symbols, in a numeric region, Pro OCR converts a letter or symbol to the number or special symbol that it most closely resembles. For example, the letter “S” in a numeric region is recognized as the numeral “5” and the letters “I” “i” “l” and the punctuation symbol “!” is recognized as the numeral “1.”

Use a numeric region whenever you want to make sure that all characters in a locate

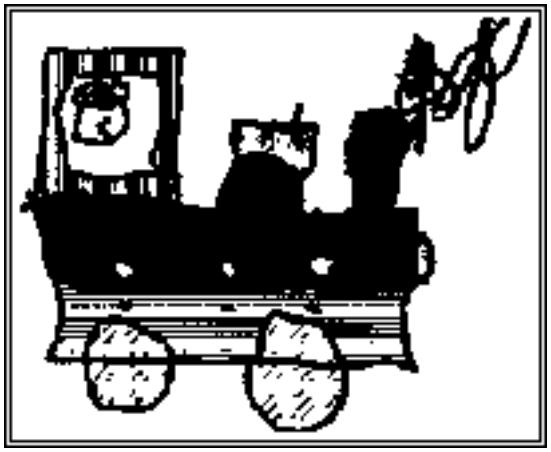
region are recognized as numbers and not mistaken for letters.

You can define numeric regions manually or with a template. Pro OCR does not define numeric regions automatically. You can also redefine a selected numeric region as any other kind of locate region using the Style menu or the Style ribbon.

Picture Regions

A picture region is a locate region that contains any kind of graphic, illustration, photograph, drawing, or picture. Pro OCR cannot recognize the contents of a picture region, but can save the image as a picture, either embedded within a document file or as a separate image file.

A picture region is enclosed in a double box:



You can create Picture regions automatically or manually, or you can predefine them with a template. To create picture regions automatically, you must select the Locate Text and Pictures option in the Gallery toolbar.

Tables

You can combine one or more text and numeric regions into a table. Use tables to help Pro OCR export tabular information correctly to other applications.

A table is enclosed in a single box. The regions it contains are shown with dimmed outlines:

Mineral	1000 tons	$ millions	% Total
Gold	2114	208.7	51.50
Asbestos	186581	51.0	9.90
Nickel	11489	108.9	22.10
Lint	65536	34.0	6.60
Chrome	561477	37.7	5.50
Big Rocks	8550110	5.8	1.10
Iron Ore	1020900	13.5	2.50
Cobalt	132	1.6	2.60

Typically, you locate a table by putting a single text or numeric region around all of the columns of the table. However, if you have a table where some columns are text and some columns are numeric, you may want to use the Make Table command. Make Table allows you to select different types of regions and then combine them into one object so that the text is exported into a tabular format, rather than columns.

Pro OCR's Locating Methods

You'll find that the locate regions that Pro OCR defines automatically are perfectly suitable for most of the pages you're recognizing. When you need more control over how Pro OCR locates a page, you can manually locate the pages yourself, or have Pro OCR automatically locate the pages and then make corrections in the text mode. You can also create a template to save the locate regions and apply the template automatically to one or more pages in one or more documents.

The section discusses Pro OCR's locating methods, including how to locate pictures as well as text, and gives you some suggestions when to use each setting.

Locating Text and Pictures

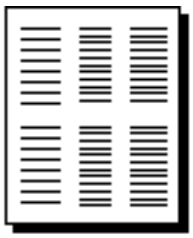
Pro OCR can automatically determine the appropriate number of columns for a page of text, using its automatic locating methods. Pro OCR has two locating methods: Multiple Columns and Single Columns Only.

In general, when you use the Multiple Columns locating method (which is the default method), text and picture regions are defined along paragraph and column boundaries. When you use Single Columns Only, Pro OCR ignores column and paragraph boundaries and defines text and picture regions that go from the left margin to the right margin of the page.

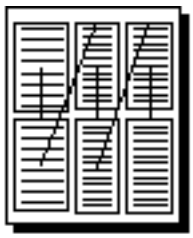
Deciding When to Use Multiple Columns or Single Column Only

Depending on the content of the page, you can organize the actual flow of the text in different ways.

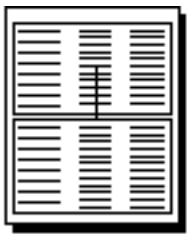
In particular, does the text flow like a newspaper article (top to bottom first, and then left to right), or does it flow like a form (left to right first, and then top to bottom):



Original input page



Flows like a newpaper article



Flows like a form or a spreadsheet

When you look at a page like this, if you understand its contents, you know which organization is the right one for the page:

- Use the Multiple Columns locating method on pages with ordinary paragraphs, pages with mixed text and graphics, and on multi-column pages of text such as in newspapers and magazines. This method also works on most pages that have tables or tabular data.

- Use Single Columns Only on pages that have side-by-side blocks of text that you want Pro OCR to read from left to right across the page. When you use Single Columns Only, Pro OCR always creates text regions that go from the left margin to the right margin of the page, regardless of the spacing of groups of words.

Because every page is different, you must experiment with using the different locating methods, so that you can understand which locating method is most appropriate for the kinds of pages that you're processing.

NOTE: You cannot define numeric regions or tables using the Multiple Columns or Single Column locating method. You can define a numeric region or a table manually. You can change a selected text region to a numeric region using the Style menu or Style ribbon, and you can group several locate regions together into a table with the Make Table command. You can also save and use a template that contains numeric regions and tables.

How to Locate Text and Picture Regions

Locating is typically done after getting a page and before recognizing. You select a locating method to tell Pro OCR how to define and order locate regions on a page. Pro OCR uses the selected locating method with automatic processing and when you click the Locate button. You can also locate regions manually. For more information, see "Defining Locate Regions Manually," later in this chapter.

To locate text or picture regions:

1. Select the locating method—**Multiple Columns, Single Columns Only,** or **Template**—from the **Locate** drop-down list in the Gallery toolbar.

2. If you select the Template locating method, select the template you want to use.

3. Select **Locate Text Only** or **Locate Text and Pictures** from the **Locate** drop-down list in the Gallery toolbar.

4. Click the **Locate** button in the Gallery toolbar.

5. If locate regions are already defined for that page, a dialog box appears that asks you if you want to discard previously defined locate regions on the page. To proceed with a new Locate step, click **Yes**.

   The document window switches to the image view, and the page zooms out to 25% of actual size. While Pro OCR is locating, the progress bar moves down the length of the page, and the status display area shows the percentage of the process completed.

   When Pro OCR finishes locating, the page appears at its previous zoom level.

Overlapping Text and Pictures

A picture region that overlaps text in a text or numeric region has no effect on the recognition of the text. The text or numeric region is recognized as if the picture region did not exist.

However, when a picture region overlaps text, the text is included as part of the picture, unless you selected the White Out Text in Pictures processing option before recognizing.

TIP: To select White Out Text, choose Options from the Tools menu, and then select White Out Text in Pictures in the Processing Options.

Locating with a Template

If the locating method you selected, Multiple Columns or Single Columns Only, doesn't work exactly as you want, you can manually create the appropriate locate regions on a page. If you have many pages that have identical layouts, such as invoices or bank statements, you can save the locate regions created on one page as a template to be used for all the pages. You can then apply that template to every page using automatic processing.

What Is a Template?

A template is simply a set of locate regions (text, numeric, or picture regions) and tables that you have saved to a file and can retrieve whenever you want to use it. After a template has been read in, you can modify the locate regions defined by it and locate the current page with the modified locate regions. You can use the modified locate regions on other pages by saving as the same template or with another template name.

Creating a Template

Create and use templates when you want to apply the same set of locate regions to many pages. You can use a template that you've already created, or you can manually define all the locate regions on one page and save them as a template which you can then apply to all subsequent pages.

When you create a template, you locate the type of document for which you want to create the template and then you manually adjust the locate regions and save the template for future use.

To create a template:

1. From the **Locate** drop-down list in the Gallery toolbar, select the Locate options that you want to apply to the document.

   For example, if you want to locate a series of single column brochures but exclude the picture, select Locate Text Only and Single Columns only.

2. Click the **Get Page** button in the Gallery toolbar.

3. Click the **Locate** button in the Gallery toolbar.

   Pro OCR locates the document.

4. Manually adjust the locate regions.

   For example, to adjust the region size, such as to exclude text, click the border of the text region and drag to include or exclude text. To delete a text region, select the border of the region and press the Delete key.

   To apply a region type, such as numeric or picture to a region, select the region border, and then click a Region Type in the Gallery toolbar.

5. To save your changes as a locate template, choose **Save Template As** from the **File** menu.

6. Type a name for the file and choose **Pro OCRTPL** as the Save as type.

   The template is now available for use with other documents. For information about using the template, see "Using a Template for Locating Regions," later in this chapter.

Using a Template for Locating Regions

Often you won't want to recognize all the information on a page. Using a template lets you select specific areas on a page that you want to recognize. Recognizing only the areas that you need to can create several kinds of savings for you:

- It cuts down on the time it takes Pro OCR to recognize a page. You might, for example, save several hours if you only recognized the bottom half of each of 200 pages.

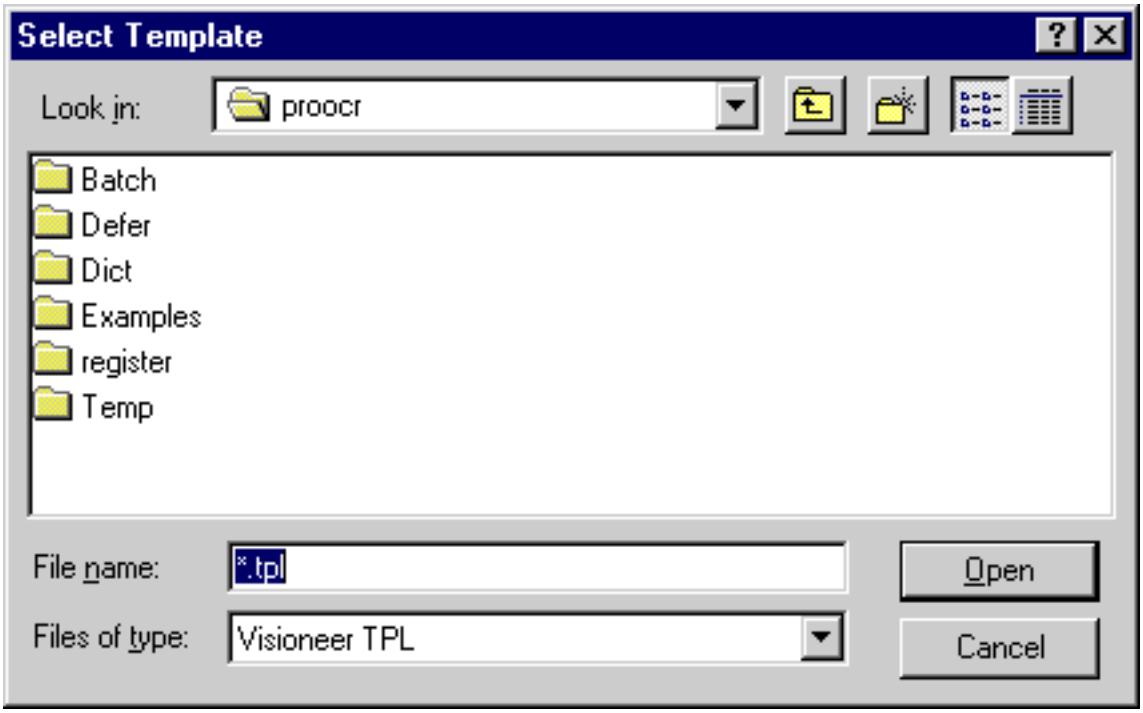
- It can save you the time of editing the saved information after it's been recognized. If you're going to discard the same portion of each page once it's exported to your word processor, it's usually more efficient to exclude that portion from being recognized.

To use a template:

1. Choose **Template** from the **Locate** drop-down list in the Gallery toolbar.

2. Choose **Select Template** from the **File** menu.

The Select Template dialog box appears.



3. Find and select the template that you want to use.

4. Click **Open**.

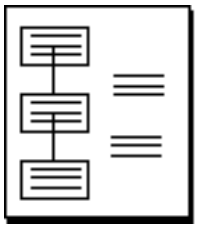
   Pro OCR displays the name of the template you selected next to Template in the Locate drop-down list.

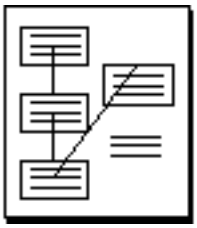
5. Get the document using the Get Page button in the Gallery toolbar.

6. Click **Auto OCR** or if you want click Locate and Recognize buttons in the Gallery toolbar to manually locate and recognize information.

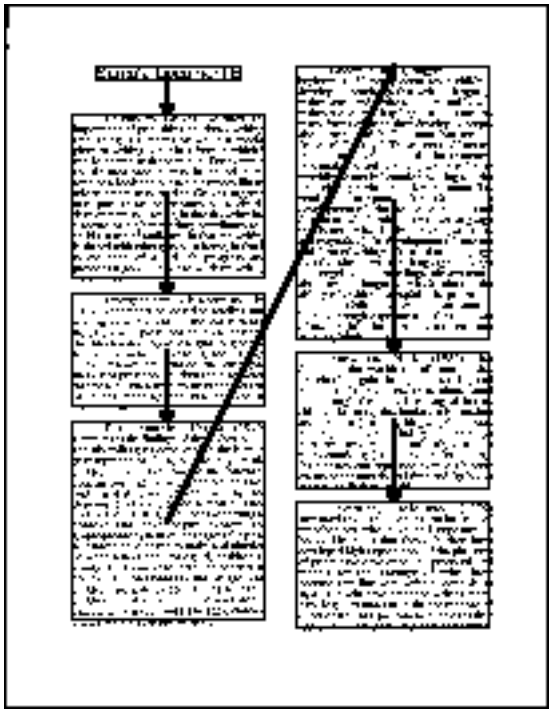
Order of Locate Regions

When a page has more than one locate region, Pro OCR automatically orders the locate regions. When you manually define locate regions, Pro OCR orders them as you create them—the first locate region you define is the first in the sequence, the second is the second, and so on. When you manually add a new locate region to a page with existing locate regions, it is added at the end of the existing sequence of locate regions on the page.


Before: Three text regions on page (arows show order)


After: Four text regions on page (arrows show new order)

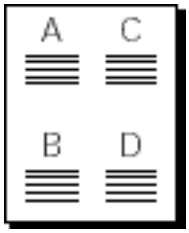
In the image view, the order of locate regions is shown by arrows from the center of one locate region to the top-center of the next locate region. This sequence tells Pro OCR in what order it should process the regions:



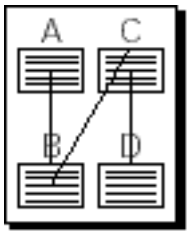
You can manually change the order of locate regions that either you or Pro OCR have defined.

The order of locate regions defines the sequence in which the information on the page is processed and output to a file. It is easiest to understand why this is important by seeing what happens to text when it is output to a word processor that has limited support for complex page layouts. In such a word processor, text regions that appear side by side in your document are output with each paragraph following the previous one in the order in which the text regions are defined. The original column margins are not preserved, and text is reflowed between the original page margins.

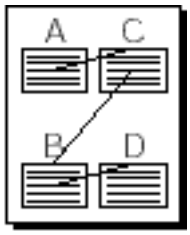
For example, if your input page has the following paragraph structure:



The order of the paragraphs (and the flow of the text) on the page might be as shown in Example 5-1 or as shown in Example 5-2:

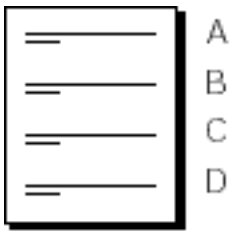


Example 5-1

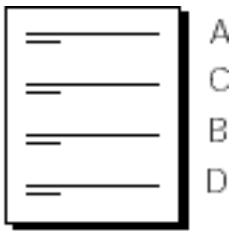


Example 5-2

When the order of text regions is defined as in Example 5-1, the text is output to the word processor as in Example 5-3. When the order of text regions is defined as in Example 5-2, the text is output to the word processor as in Example 5-4:



Example 5-3



Example 5-4

When you automatically process (with Auto Start, Finish Processing, Process Deferred Jobs), or single-step Locate with the manual locating method, Pro OCR automatically orders all locate regions. If the assigned order does not correspond with the way the text flows on a page, you can reorder the locate regions so that they'll be output in the correct sequence. For more information, see "Reordering Locate Regions," later in this chapter.

Examples of Locating Documents

Some documents require special care when processing. The following examples show you how to configure Pro OCR to process these documents properly.

Processing Resumes

For Pro OCR to process resumes and legal documents properly, select **Single Columns Only** from the **Locate** drop-down list in the Gallery toolbar.

Resumes often contain formatting elements that can be difficult for an OCR program to interpret, such as numerous indentations, bulleted items, and a wide mixture of both justified and centered text.

NOTE: if you have already located the page using a setting other than Single Column Only and want like to re-locate the page, simply change the setting and click the Locate button.

Processing Legal Documents

Like resumes, legal documents often contain formatting elements that can be difficult for an OCR program to interpret properly. Usually legal documents, such as court papers, contain case or document information at the top or top right of the page, numbers along the left side of the page, and a wide mixture of indented and centered text.

Sample Document D, which is included with the software, shows the formatting elements typical of a legal document. You can use Sample Document D to test Pro OCR's legal document handling abilities.

Processing Faxed Documents

Pro OCR's features make fax recognition easier and more accurate than ever before:

- The Page Image Rotation commands allow you to correct the orientation of upside-down or sideways faxes without having to use another imaging or fax program.

- The Degraded or Fax Quality option in the Recognize drop-down list helps clean up any dirty or "noisy" faxed pages which are often the result of poor phone connections when faxing.

About Columns, Locate Regions, and Output File Formats

Pro OCR preserves virtually all page layout and text flow information in the documents it processes. However, when you save to a specific word processor format, Pro OCR preserves only as much of this layout information as the particular application format is designed to use. Some applications can use (interpret, display, and print) more of this kind of information than others.

NOTE: You can also specify which page layout and text flow information is saved by opening the Save As Options dialog and changing its settings.

Some word processors have extensive support for complex document layouts, while others provide only limited support.

Defining Locate Regions Manually

For most pages, you'll locate automatically as part of automatic processing, Finish Processing, Process Deferred Jobs, or on a page-by-page basis using the Locate button. Sometimes, however, you'll want to define locate regions manually before recognizing. Manually locate when you want to:

- Recognize and save only some of the text on a page.
- Save text in a different order than it's automatically located.
- Resize an existing locate region.
- Delete a previously defined locate region.
- Add a new locate region to locate regions that are already defined.
- Define a numeric region.
- Redefine one type of locate region as another type of locate region (for example, redefine a text region as a numeric region).

- Create tables.

You can use manually located regions to create a template, just as you can create templates from automatically located text regions.

As with all other locating procedures, you can only manually define locate regions in the image view by selecting Image from the View menu.

When you manually locate, you specify the size and extent of one or more locate regions on the page using the mouse. If there is more than one locate region on a page, you also determine the order in which they are processed during Recognizing and output to a file.

After you manually locate a page, you can immediately do Recognize using the locate regions you've defined. You can also save the locate regions to a template. You can then use that template to apply the same locate regions to other pages in this and other documents. You can also locate additional pages of the document and then use Finish Processing, or save the document in the Pro OCR Deferred format to be processed later using Process Deferred Jobs.

To manually create a new locate region:

1. Move the pointer outside of any existing locate regions, in the image view.

   Whenever the pointer is outside any existing locate regions, it turns into a cross hair pointer. When it is within an existing locate region, it is the standard arrow pointer. When it is over a sizing handle, it is a resizing pointer.

   NOTE: You must start a new locate region outside of any existing locate region.

2. Drag the cursor over the are that you want to locate.

3. Select the type of region from the Gallery toolbar based on the following information. Use the following table to create the region you want:

Icon	Region	Do this...
	Text region	Drag the cross hair pointer across the page image.

	Numeric region	Hold down the Ctrl key as you drag the cross hair pointer across the page image.
	Picture region	Hold down the Ctrl + Shift keys, as you drag the cross hair pointer across the page image.

As you drag, a box is drawn from the corner where you started to the cross hair of the pointer. When the box encloses the desired text, release the mouse button.

A box is then displayed with sizing handles in each corner and at the center of each side. A text region is enclosed in a box drawn with a single solid line. A numeric region is enclosed in a box drawn with a double solid line with dots between the lines. A picture region is enclosed in a box drawn with a double solid line.

If there are any locate regions already on the page, an arrow is drawn from the center of the existing locate region on the page that is last in the sequence, to the top center of the new locate region you've just created.

Tips When Creating Locate Regions

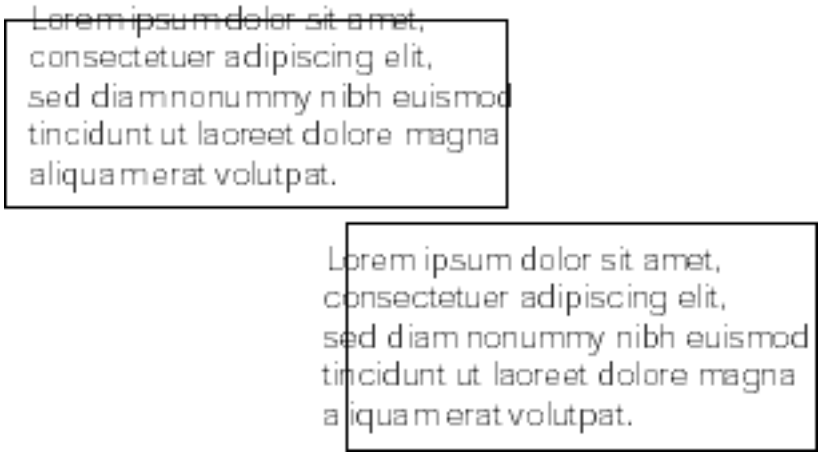
The following tips may help you when creating a new locate region:

- Locate regions are always ordered in sequence. When you define a new a locate region, it is placed at the end of the existing sequence of locate regions.
- A locate region cannot be created smaller than 21 image pixels on a side. If the mouse button is released before a locate region is at least this big, no locate region is defined.
- A locate region cannot extend beyond the edges of the page image. If a new locate region extends beyond the edge of the window during resizing, the window automatically scrolls up to the edge of the page.
- You can manually resize locate regions—both the ones that Pro OCR creates automatically and the ones you create manually. Locate regions are always resized as a rectangle.

Overlapping Locate Regions and Skewed Text

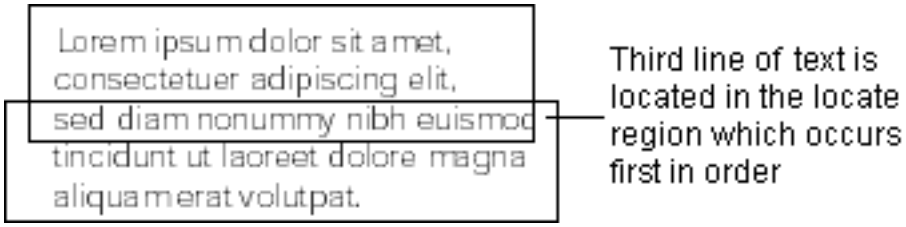
When you manually create text or numeric regions, you should be aware of the following constraints.

If a text or numeric region cuts through a character, only the part of the character that is within the region is located. When these cutoff characters are recognized, most of them will be illegible. This can happen when the characters are cut off at the sides or when they're cut off at the top or bottom:

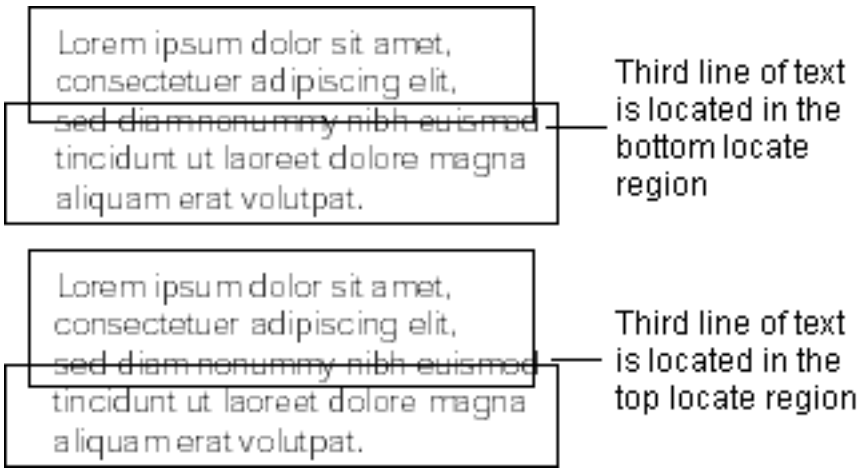

Lorem ipsum dolor sit amet,
consectetuer adipiscing elit,
sed diam nonummy nibh euismod
tincidunt ut laoreet dolore magna
aliquam erat volutpat.
Lorem ipsum dolor sit amet,
consectetuer adipiscing elit,
sed diam nonummy nibh euismod
tincidunt ut laoreet dolore magna
aliquam erat volutpat.

NOTE: If the characters along the edges of a text or numeric region you've defined manually are illegible, go back to the image view and check to make sure that the text or numeric region does not cut off any of the edges of the text image.

If a line of text or numbers is enclosed within more than one text or numeric region, it is located only once. If the line is fully enclosed within both regions, it is located within the text or numeric region that is ordered first in the sequence:

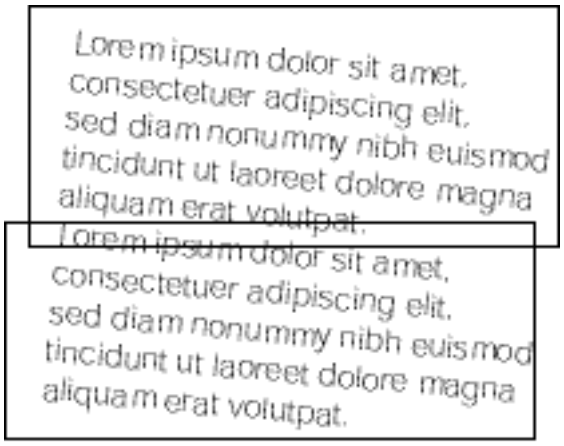

Lorem ipsum dolor sit amet,
consectetuer adipiscing elit,
sed diam nonummy nibh euismod
tincidunt ut laoreet dolore magna
aliquam erat volutpat.
Third line of text is
located in the locate
region which occurs
first in order

If the line is fully enclosed within one region and only partially enclosed within another region, the line is located within the text or numeric region that fully encloses it:


Lorem ipsum dolor sit amet,
consectetuer adipiscing elit,
sed diam nonummy nibh euismod
tincidunt ut laoreet dolore magna
aliquam erat volutpat.
Third line of text is located in the bottom locate region
Lorem ipsum dolor sit amet,
consectetuer adipiscing elit,
sed diam nonummy nibh euismod
tincidunt ut laoreet dolore magna
aliquam erat volutpat.
Third line of text is located in the top locate region

These constraints are especially important when pages are skewed (read in crooked). Because locate regions are defined by rectangles that are square to the screen, when you have skewed text in a document, you may have to overlap text or numeric regions in order to not cut off any lines and get all the text into the appropriate region. When this happens, if the locate regions are close together, sometimes a line ends up being contained within more than one text or numeric region.

The lines that are contained in both regions are only recognized if fully enclosed by at least one text or numeric region. In this example, the fifth line is located in the top text region, and the sixth line is located in the bottom text region:


Lorem ipsum dolor sit amet,
consectetuer adipiscing elit,
sed diam nonummy nibh euismod
tincidunt ut laoreet dolore magna
aliquam erat volutpat.
Lorem ipsum dolor sit amet,
consectetuer adipiscing elit,
sed diam nonummy nibh euismod
tincidunt ut laoreet dolore magna
aliquam erat volutpat.

You need to make sure that the line that is contained in the desired text or numeric region is fully enclosed by it. Otherwise, any characters that are cut off by the locate region will be illegible.

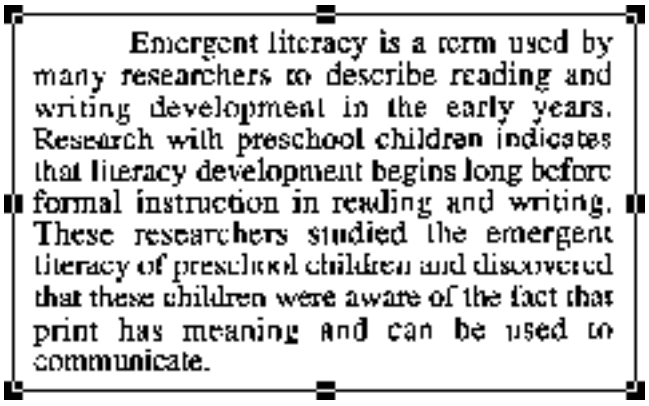
If all the text in your document is skewed the same way, you may use the Straighten Skewed Images processing option to straighten the page image when it is read in. This will usually eliminate the problem of overlapping regions. However, this processing option will slow Pro OCR down.

For more information about Straighten Skewed Images, "Setting Scanning Options," in Chapter 3.

Selecting and Deselecting Locate Regions

You can only select locate regions in the image view. You select a locate region to change its kind, delete it, or resize it.

When any locate region is selected, sizing handles appear:


Emergent literacy is a term used by many researchers to describe reading and writing development in the early years. Research with preschool children indicates that literacy development begins long before formal instruction in reading and writing. These researchers studied the emergent literacy of preschool children and discovered that these children were aware of the fact that print has meaning and can be used to communicate.

To select a single locate region:

1. Move the pointer over the locate region.

   When the pointer is over a locate region, it is the standard arrow pointer.

2. Click anywhere in the locate region.

   When you select a locate region when other locate regions are selected, the previously selected locate regions are deselected.

To select more than one locate region at a time:

1. Select a single locate region as above.

2. Move the pointer over another locate region and shift-click.

   The selected regions are displayed with a thick border.

3. Repeat Step 2 for each additional locate region you want to select.

To deselect one or more locate regions while keeping the rest selected:

- Move the pointer over a selected locate region and shift-click.

  The locate region you clicked in is deselected, but all other selected locate regions stay selected.

  Repeat this step for each locate region you want to deselect.

To select all locate regions:

- Choose **Select All** from the **Edit** menu.

  All locate regions defined for the page are selected.

To deselect all locate regions:

- Click anywhere outside of any locate regions.

  All selected locate regions are deselected. For more information, see "Resizing a Locate Region," later in this chapter.

Changing the Kind of a Locate Region

You can only redefine a locate region in the image view. You redefine a locate region to change it to a different type of locate region. You can change the kind of a locate region to any other kind of locate region.

To redefine a locate region as any other type of locate region:

1. Select the locate region to be redefined.

2. Choose **Text, Numeric,** or **Picture** from the **Style** menu or **Style** ribbon.

The selected locate region is changed to the specified type of locate region.

Deleting a Locate Region

You can only delete a locate region in the image view. You delete a locate region when you don't want the image in that locate region to be processed, or when you

want to define a different locate region that includes the image in that region.

NOTE: Only the defined locate region is deleted, not the underlying image. The underlying image never gets deleted.

To delete a locate region:

1. Select the locate region to be deleted.

2. Press **Delete** to remove the selected locate region, or choose **Clear** from the **Edit** menu to remove the selected locate region.

   The box around the image of the text disappears.

   If there were two locate regions on the page, the arrow connecting the remaining locate region to the deleted locate region disappears, and there is only one locate region on the page. If there were more than two locate regions on the page, the deleted locate region disappears, and the order of the remaining locate regions remains the same.

Resizing a Locate Region

You can only resize a locate region in the image view.

To resize a locate region:

1. Select the locate region.

   The locate region's sizing handles are shown, if it is the only locate region selected.

2. Move the pointer over one of the locate region's sizing handles.

   Whenever the pointer is over a corner sizing handle, it turns into the four-arrow pointer, and whenever the pointer is over a side sizing handle, it turns into a vertical or horizontal two-arrow pointer.

3. Click the sizing handle and hold the mouse button down while you drag the dotted outline of the locate region to its new size.

4. Release the mouse button.

The same conditions on the size, overlap, containment, and extent of locate regions apply to a resized locate region as to a newly created locate region.

Reordering Locate Regions

You can only reorder locate regions in the image view. Whenever you create a locate region, Pro OCR automatically links all locate regions on the page in sequence. You reorder locate regions when you want to change the automatic sequence in which they are processed and output. You can reorder any locate regions.

To reorder locate regions:

1. Move the pointer over a locate region.

   The pointer becomes the standard arrow pointer.

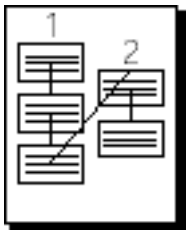
2. Click in the locate region (make sure you don't click on a sizing handle).

3. Hold down the mouse button and drag the pointer into the locate region to which you want to relink, and then release the mouse button.

   The arrow originally connecting the locate region you're relinking to disappears, and the new arrow connects the preceding locate region to the newly linked locate region. The remaining regions are reordered as close to the original order as possible.

The following example shows you how reordering works in one specific case.

The page has been processed using the Normal locating method and has the following locate regions, linked in the order shown:



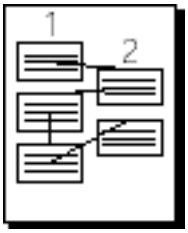
You want to change it so that locate region #2 is in order after locate region #1.

To relink (an example):

1. Move the pointer over locate region #1 and press and hold down the mouse button.

2. Drag the pointer into locate region #2, then release the mouse button.

   The arrow originally leading into locate region #2 disappears, and a new arrow connects locate region #1 to locate region #2:



© Copyright 1998 Visioneer, Inc. Reach us at www.visioneer.com.

Chapter 5

Setting Recognize Options and Proofing a Recognized Document

When you recognize a document you convert an image into editable text. You can then proof and edit the text. This chapter tells you how to:

- Select the type quality option for recognizing.
- Select display options, including the fonts that Pro OCR uses when recognizing text and when displaying the recognized document, suspect threshold level, and illegible character symbol.
- Select proofing options.
- Use the proof command to proof a recognized document.
- View and edit a recognized document in the text mode.
- View a summary of errors for a recognized document.

NOTE: You can recognize text automatically by using Auto OCR or you can recognize text in a single step. For more information about Auto OCR, see the examples in Chapter 2.

Selecting Type Quality Options

Use the type quality options to tell Pro OCR whether you're recognizing laser printed text, dot matrix text or degraded or fax quality text. You can select these options in the Recognize drop-down list in the Gallery toolbar.

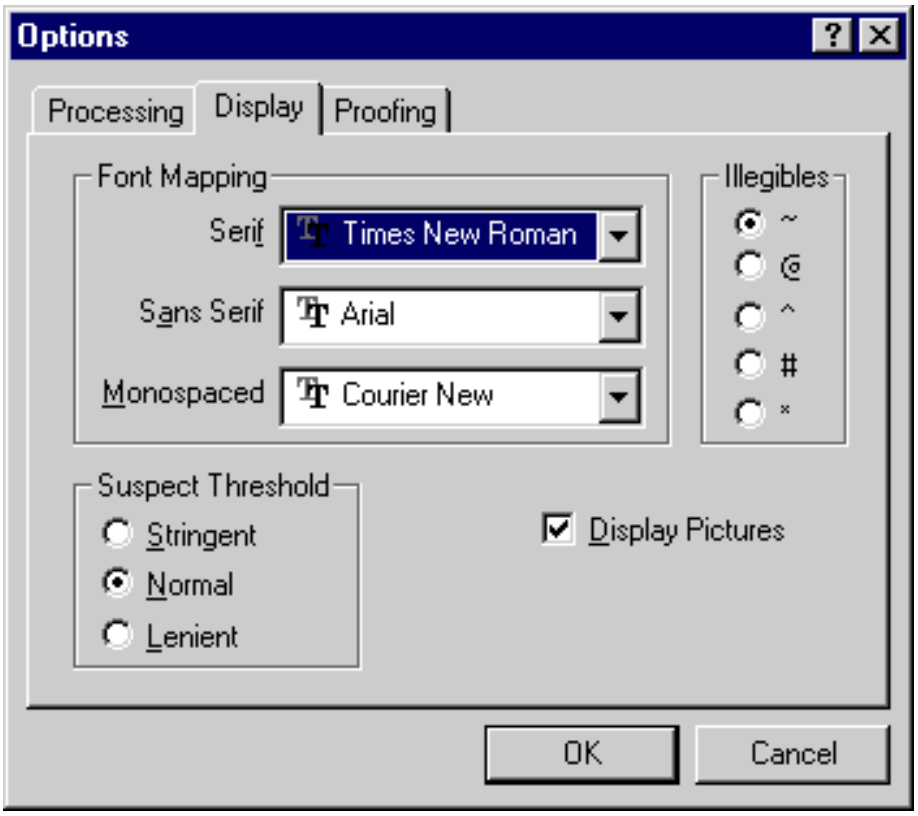
For most documents, you'll select Letter Quality. Select Dot Matrix Quality only when the characters in the input document are in monospaced type and made up of dots that are not touching. Select Degraded or Fax Quality when you are recognizing a document with less than optimum text.

To select the type quality for recognizing:

- Select a type quality from the **Recognize** drop-down list in the Gallery toolbar.

Selecting Display Options

Use the Options dialog box to select options that tell Pro OCR how to recognize a document and display the results.



In the Options dialog box, you can select Display options that:

- Select the fonts with which you want your document displayed and exported
- Set the suspect character threshold
- Specify the illegible character symbol
- Select whether to display pictures while editing

Setting the Suspect Character Threshold

Pro OCR recognizes and identifies over 2,000 typefaces. It can make correct judgments about character identity even when a character isn't absolutely clear. However, sometimes Pro OCR cannot identify with certainty what a particular character is, and other times Pro OCR cannot identify a character at all.

To handle cases like these, Pro OCR tries to assign the correct character to a questionable character image. It also keeps track of when it has done this so that you can inspect and verify the assignment, if you choose.

Use the suspect character threshold to tell Pro OCR which flagged suspect characters to highlight. It does not affect the accuracy of the recognition. Pro OCR is always as accurate as it can be during the recognition process, regardless of the selected suspect threshold setting.

NOTE: The suspect threshold doesn't change how Pro OCR decides on or assigns the identity of a character. Thus, changing the suspect threshold doesn't change how many characters in the document Pro OCR is sure about, but only how it displays those characters to you.

You can select among stringent, normal, and lenient thresholds. After you recognize a document, the number of suspect and illegible characters highlighted on the current page is shown in the Status Display area.

Each time you clear a suspect or illegible character on a page, the number of suspect or illegible characters is decreased by one.

TIP: If you want to see the number of suspect and illegible characters in the entire document, choose Properties from the File menu. For more information, see "Displaying a Summary of Recognized Errors"later in this chapter.

To set the suspect character threshold:

1. Choose **Options** from the **Tools** menu.

   The Options dialog box appears.

2. Click the **Display** tab.

   The Display options appear.

3. Select one of the following threshold levels:

Option	Does this...
Stringent Suspect Threshold	Identifies ALL suspect characters. Use the stringent setting when it is important that you know about all possible mistaken identifications, or when using dictionaries will not aid in identification. For example, use Stringent when you recognize tables of numbers, documents with a lot of proper names, and whenever you need to check the recognition results very carefully.
Normal suspect threshold	Identifies only suspect characters of which it is somewhat uncertain. Typically, Pro OCR highlights some suspect characters at the Normal threshold. Use the normal setting with ordinary (clean, clear, typeset) documents, where accuracy is important but not critical, and when most of the words in the document are likely to be found in the dictionaries.
Lenient suspect threshold	Identifies only suspect characters of which it is very uncertain. Typically, Pro OCR highlights very few suspect characters at the Lenient threshold. Use the lenient setting for documents containing fonts that you know from experience have been recognized accurately in the past or when you're less concerned with proofing your document.

4. Click **OK**.

   When you return to the document, it appears with the new settings. The last set display options are remembered when you run Pro OCR again.

Setting the Illegibles Character Symbol

You select the illegibles character symbol to tell Pro OCR how to display any illegible characters it finds. When you use Proof with the Illegibles proofing option selected, it finds illegible characters so that you can review and edit them.

The choices for the illegible character symbol include:

~ @ ^ # *

The preset symbol is “~”. Every illegible character is represented by the same selected illegible character symbol.

Choose a symbol that you otherwise don’t expect to have in your document, so that when you search for it you will only find the illegible characters. For example, you would not use the “#” sign if your document has tables with the “#” sign in them.

All uncleared illegible characters in the document always appear, no matter what suspect threshold you’ve selected.

NOTE: If you find a lot of suspect or illegible characters in any document, make sure that you insert the pages into the scanner straight and in the correct orientation for the scanner and the page orientation you’ve selected. If necessary, make sure the “Straighten Skewed Images” processing option is selected. Also, make sure that the brightness level for your scanner is set to an appropriate setting. Additionally, make sure that Draft Quality is not selected, unless you are scanning draft quality dot matrix text.

To set the illegible character threshold:

1. Choose **Options** from the **Tools** menu.

   The Options dialog box appears.

2. Click the **Display** tab.

   The Display options appear.

3. Select an illegible character symbol in Illegibles.

4. Click **OK**.

   When you return to the document, it’s displayed with the new settings.

The last set display options are remembered when you run Pro OCR again.

Selecting a Display Font

Although Pro OCR recognizes and identifies over 2000 typefaces, it is unlikely that

you'll always have the same fonts installed in your system as the fonts identified in the input document.

To maintain as much similarity to the input document as possible, Pro OCR maps any identified fonts to three user-selectable fonts installed in your system: one monospaced font, one serif font, and one sans serif font. These fonts are used to display the recognized text of the input document for screen display and for output to the various file formats. They may be changed at any time before or after recognition. Any fonts installed in your system may be selected.

If there is more than one serif, more than one sans serif, or more than one monospaced font in the input document, you can still choose only one of each. All serif fonts are mapped to the serif font you specify, all sans serif fonts are mapped to the sans serif font you specify, and all monospaced fonts are mapped to the monospaced font you specify.

NOTE: If you are not running Windows with TrueType™, we recommend that you install a type display manager, such as Adobe Type Manager™ (ATM), before you use Pro OCR. This will make the type on your screen easier to read when you're viewing recognized text.

To select the display font:

1. Choose **Options** from the **Tools** menu.

   The Options dialog box appears.

2. Click the **Display** tab.

   The Display options appear.

3. Select the fonts you want to use from the Serif, Sans Serif and Monospaced drop-down lists.

   The font name you select appears in the appropriate box.

4. Click **OK**.

   When you return to the document, it's displayed with the new settings. The last set display options are remembered when you run Pro OCR again.

If you change the settings for Font Mapping while a document is being displayed and then return to the document, the display is updated to show the new fonts.

TIP: You can display and proof a document using one display font or set of display fonts, and then change the settings to save the document with other fonts. This is one reason for saving a separate settings file.

Indicating Whether Pictures Appear During Text View

Use the Display Pictures option to tell Pro OCR whether to display pictures in the text view. If you deselect this option, a blank box appears in place of the pictures in text view.

To select whether to display pictures:

1. Choose **Options** from the **Tools** menu.

   The Options dialog box appears.

2. Click the **Display** tab.

   The Display options appear.

3. To display pictures in text view, select the **Display Pictures** checkbox.

   To prevent pictures from appearing in text view, deselect the Display Pictures checkbox.

4. Click **OK**.

   When you return to the document, it appears with the new setting. The last set display options are remembered when you run Pro OCR again.

Recognizing a Single Page

You can use the single-step Recognize operation after locate regions have been defined for a page. When you use the single-step Recognize operation, you recognize only one page at a time. When only some pages in a file have been located, you can use this command to recognize any pages that are located and manually skip any pages that have not yet been located.

To recognize a page:

1. If you haven't already done so, locate the regions that you want to use for the Recognize step.

You can locate region automatically using Locate with any of the locate settings, or they can be located manually. For more information about locating, see Chapter 4, “Locating Text and Graphics.”

2. Select either **Letter Quality, Dot Matrix Quality,** or **Degraded or Fax Quality** from the **Recognize** drop-down list in the Gallery toolbar.

   For most documents, you’ll select Letter Quality. Select Dot Matrix Quality only when the characters in the input document are in monospaced type and made up of dots that are not touching. When you select Draft Quality, Pro OCR adjusts its recognition process to accurately recognize characters made up of dots that are not touching.

3. Click the **Recognize** button in the Gallery toolbar.

   If text has already been recognized for that page, a dialog box appears that asks you if you want to discard previously recognized text on the page. To proceed with recognition, click **Yes**.

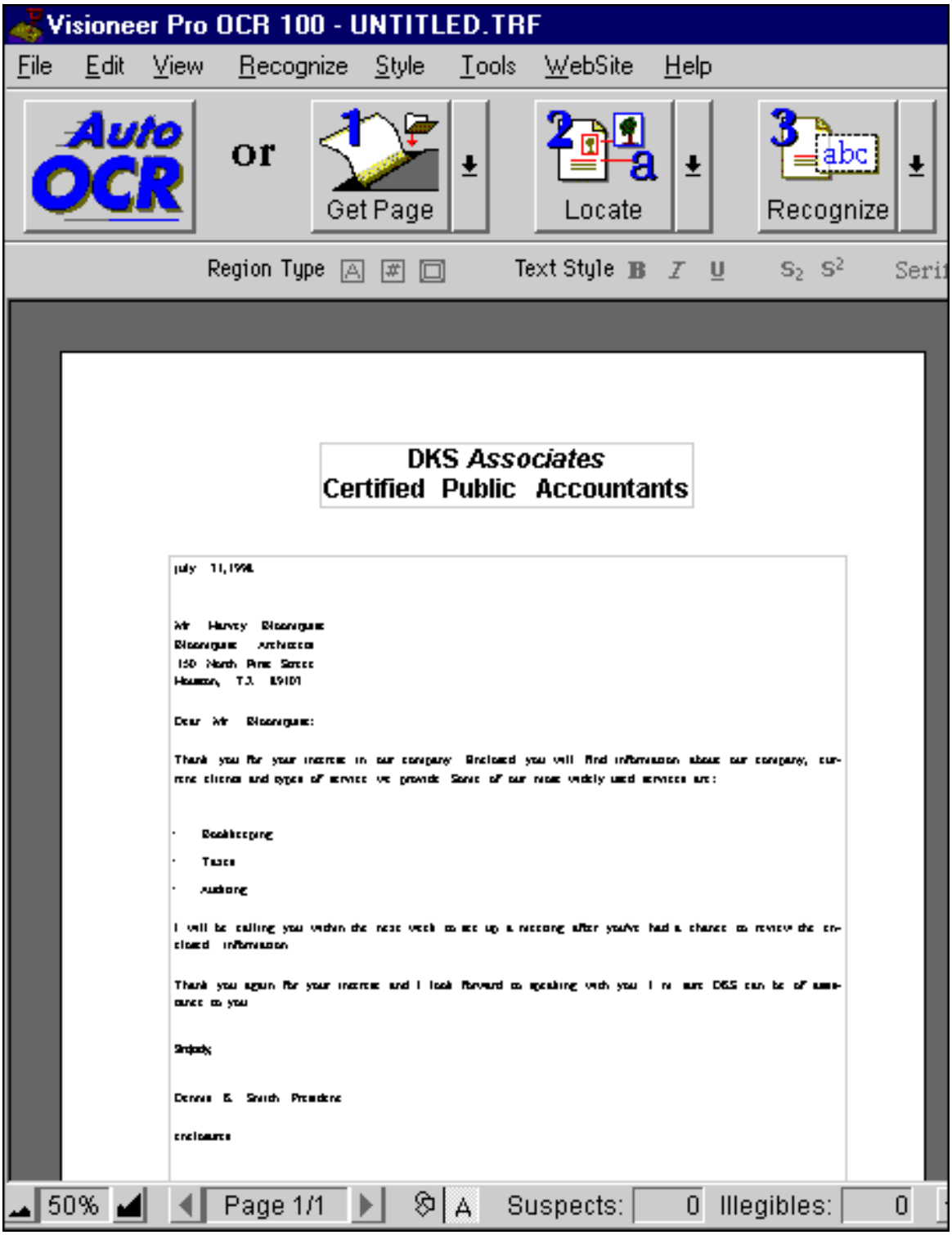
   The document window is switched to the image view, and the page is temporarily zoomed out to 25% of actual size. While Pro OCR is recognizing, the progress bar moves down the length of the page, and the status display area shows the percentage of the process completed.

   When Pro OCR finishes recognizing, the page is displayed at its previous zoom level and the document window is switched to the text view.

NOTE: If you stop Recognize in progress, text that has been recognized is discarded. The page has the located regions but no text has been recognized.

Working with Recognized Pages in Text view

When Pro OCR finishes recognizing, it displays the recognized page in the text view, highlights all characters it flagged as illegible, and highlights flagged suspect characters according to the Proofing options that are selected.



Setting the Zoom Levels

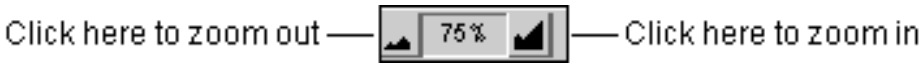
The zoom controls are active in both the image view and the text view. Use them to change between zoom levels.

You cannot zoom in closer than the pixel-for-pixel level (in the image view), or 400% (in the text view), or zoom out farther away than 25% in either view. When you're at the maximum zoom level, the zoom in control is dimmed. When you're at the

minimum zoom level, the zoom out control is dimmed.

To zoom in or out:

- Click the **Zoom In** or **Zoom Out** icon on the Status bar.



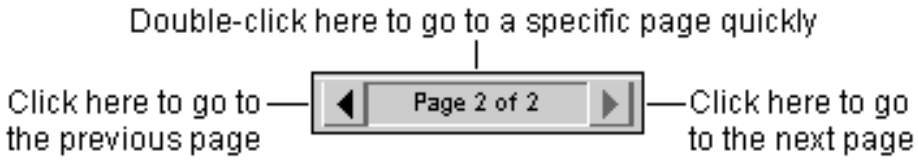
Selecting a Page to Display

The page controls are available in both the image view and the text view. The page number box in between the page controls tells you what page of the open document is being displayed and how many pages there are in the document. You can display pages sequentially or skip forward or backward to a specific page.

To move forward or backward one page:

- If the document has more than one page, click the arrows to change pages.

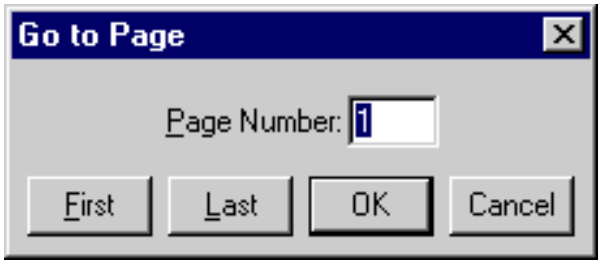
  If you're on the first page, the previous page arrow is dimmed. If you're on the last page, the next page arrow is dimmed.


Double-click here to go to a specific page quickly
Click here to go to the previous page
Page 2 of 2
Click here to go to the next page

To display a specific page:

1. Double-click the page number box that appears between the two arrows, or choose **Go to Page** from the **View** menu.

   The following dialog box appears:



   The current page number is displayed and highlighted.

2. Type the number of the page you want to go to and click **OK**, or click the **First** or **Last** button to go to the first or last page.

The requested page appears. The page number box changes to the new page number.

Selecting Text or Image View

The View controls are current in both the image view and the text view. Use them to change between the image view and the text view. Pro OCR highlights the selected button to indicate which view you're currently in.

To change views:

- Click the **Image View** icon or the **Text View** icon.

Proofing

After a document is recognized, it appears in the text view. In this view, you can use the Proof command to proof and edit the document.

Pro OCR keeps track of any characters that it couldn't recognize (illegible characters), and track of characters that it wasn't certain it had recognized correctly (suspect characters), and highlights them. You can use the Proof command to:

- Systematically inspect recognized text and edit it if necessary.
- Search for misspelled words, numbers, punctuation, symbols, and alphanumeric words.
- Add any specialized words in the document to a user dictionary, at any time.

You can proof and edit each line of displayed text on a line-by-line basis or by using Proof with the "Whole Lines" proofing option selected. You can also search for and replace words, one by one or all at once.

The following sections tell you how to:

- Select Proofing options
- Start proofing

- Edit a document

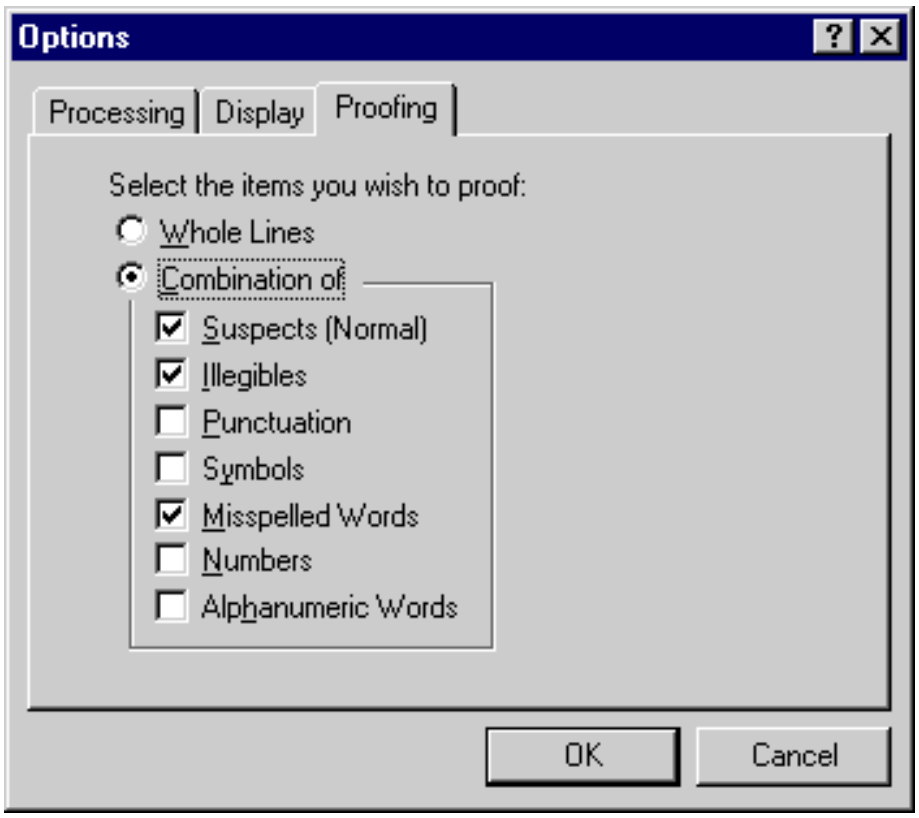
Selecting Proofing Options

Set Pro OCR Proof options to indicate if you want to proof whole lines and what combinations of words and punctuation you want to proof.

To select Proofing options:

1. Click the down arrow next to the Proof button in the Gallery toolbar.

   The Options dialog box appears with the Proofing options displayed.



2. Select one of these options.

   - **Whole Lines.** Proofs the entire document one line at a time. Each time you choose Proof, the insertion point moves to the start of the next line, and the On-Screen Verifier is displayed. Use this option if your proofing style is to quickly scan each line for any errors.

   - **Combination Of.** Proofs a combination of characters. Select this option and then select whichever options you want to combine. Pro OCR moves through the document one specified character or word at a time.

3. (Optional) If you select Combination Of, select any of the following options:

   - **Proof Suspect and Illegibles.** Pro OCR selects each suspect or illegible character as it is encountered. Note that Pro OCR uses the selected suspect threshold display option to decide which characters are suspect.

   - **Proofing Punctuation and Symbols.** Pro OCR searches for and selects each punctuation mark or symbol as it is encountered. The set of punctuation characters includes:

     ; . , : - ~ ? ! ' " ´

     { } [ ] ( ) < > ‘ ’

     The set of symbol characters includes:

     © ® ™ @ ¶ § # $ ¢ ° £ ¥ &

     * = ± ÷ + x - \ / % ^ | ¼ ½

   - **Proofing misspelled words.** Pro OCR checks each word it encounters to see if it is in the General dictionary, the current user dictionary, or any other dictionaries that are installed in the Dictionaries directory.

   - **Proofing Numbers and Alphanumeric Words.** Pro OCR searches for and selects each number or alphanumeric word as it is encountered. A number is a word consisting of numeric characters (0–9), and the following characters: + - , . Numbers are bounded by tabs and spaces. An alphanumeric word is a word consisting of any of the alphabetic and numeric characters (A–Z, a–z, 0–9), excluding punctuation and other symbol characters. Alphanumeric words are bounded by tabs and spaces.

4. Click **OK**.

NOTE: When Proof selects a word while the Misspelled Word proofing option is selected, it’s not necessarily misspelled—it might be that the word isn’t in your current user dictionary or any of the dictionaries in the Dictionaries directory.

Proofing a Document

If a misspelled word is encountered that contains a suspect character or an illegible

character, the suspect or illegible character is visited and selected first, and the next time you use Proof, the same word is selected again.

To proof a document:

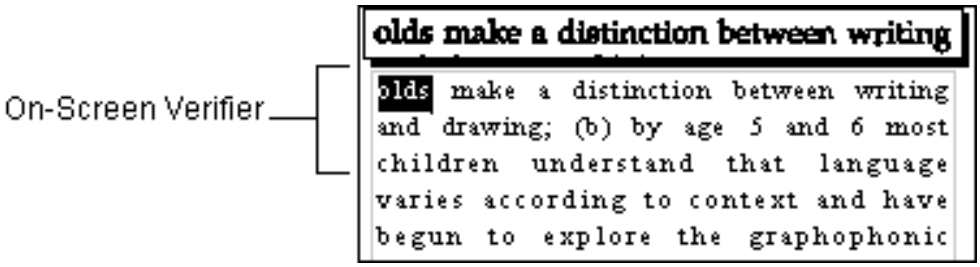
1. In the text view, click the **Proof** button, or choose **Proof** from the **Recognize** menu.

   TIP: You can also press the Tab key to start the proof.

   Pro OCR starts at the current insertion point, if there is one. Otherwise, it starts at the top of the current page.

   Depending on which proofing options you've selected, the next marked suspect or illegible character, or the next specified word or character, is found and selected. The document is scrolled so that the character or word is in view.

   If the image of the page exists, the On-Screen Verifier (shows you the actual image of the corresponding portion of the document.


On-Screen Verifier
olds make a distinction between writing
olds make a distinction between writing and drawing; (b) by age 5 and 6 most children understand that language varies according to context and have begun to explore the graphophonic

2. Switch between the currently selected zoom level and the pixel-for-pixel zoom level by clicking anywhere in the pop-up window.

   NOTE: To turn the On-Screen Verifier on or off, choose Proofing Verifier from the View menu.

3. Inspect the selected character and, if necessary, type in the text correction.

   You can use editing commands, such as Cut, Copy, Paste, Clear, Find &

Replace to make changes.

TIP: If the selected text is misspelled, and you expect to find further instances of the word in this document, don't edit it. Instead, use Find & Replace. The selected word is displayed as the Find text. You can type in the correct spelling in the Replace text box and change all instances of the word, either with Replace or Replace All.

NOTE: When you add a word to the current user dictionary, it's available immediately. Proof will not find further instances of the word. For more information about the User dictionary, see "Using Dictionaries in Pro OCR" later in this chapter.

4. Repeat Steps 1 and 2 for each specified character or word proofing option, until no more errors are found.

   If you did not start at the beginning of the document, a message asks if you want to continue from the beginning of the document.

5. Click **OK** to return to the beginning and check the rest of the document.

NOTE: If you've displayed the page with, for example, the Lenient suspect threshold, and you've cleared all suspect and illegible characters, you can change the Suspect Threshold display option to Normal or Stringent, and choose Proof again.

Reviewing and Editing Text in the Text View

You can view and edit text in the text view. Pro OCR highlights suspect characters and illegible characters and marks them with a specified Illegible Character symbol. In text view, you can:

- Search for and replace words, using Find & Replace and Find Again.
- Add words to the current user dictionary by choosing Add to User Dictionary from the Tools menu.
- Use the View menu or the zoom controls in the Status bar to display the page at 25 to 400 percent.
- Switch between displaying pages in text view and image view.

To display text view:

- Click the text view button in the Status bar, or by choosing **Text** from the **View** menu.

To edit text within a line:

1. Move the pointer over the text line.

   The pointer indicates the text selection.

2. Click anywhere within the line.

   The line becomes active for editing. The blinking vertical bar cursor indicates where the insertion point is on the line.

   Clicking anywhere outside the active line deactivates it. If the click is in another text line, that line becomes active.

3. Edit the line.

4. To edit a different line, repeat steps 1 and 2 or use the arrow keys to move to a new line.

NOTE: When you've selected text manually, the On-Screen Verifier is not automatically displayed. You can display it by choosing Proofing Verifier from the View menu.

Standard Text Editing Operations

The following standard text editing operations are available to edit text within a line in the text view.

To	Do this...
Select text	Click and drag.
Extend the selection one word to the left or right	Ctrl-Shift-left/right arrow
Extend the selection to the beginning or the end of the line	Shift-Home/End
Select a word	Double-click the word
Select contents of an entire line	Triple-click a word in the line.

Cut, copy, paste, clear	Use keyboard equivalents or click the right mouse button
to select or deselect characters one at a time.	Hold down the Shift key while using the arrow keys

Lines don't wrap when more characters are added to a line. Instead, text is squeezed into the existing line, squeezing the space between characters and words proportionally and overlapping them if necessary.

Lines don't rewrap when characters and words are removed from a line. Instead, the text is stretched to fill the available line length, with space between characters and words extended proportionally.

In either case, when the text is saved in a file format (for example, in a word processor file format) that supports text wrap, the text can reformat—line breaks might not be preserved and might be rewrapped. (Carriage returns will be preserved when saving to the Text With Line Breaks format or to a spreadsheet format.)

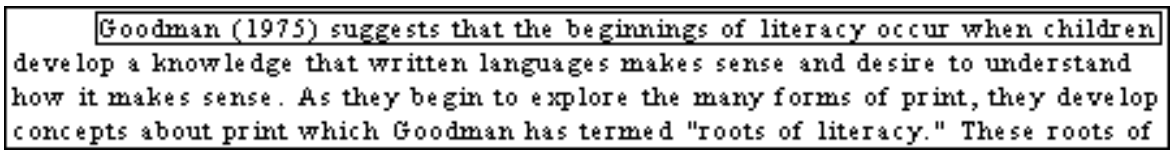
To select a single text line:

1. Move the pointer over the text line.

2. Hold down the **Ctrl** key and click anywhere in the text line (**Ctrl-click**).

   (When the Ctrl key is held down, the pointer becomes the standard arrow pointer.)

   A box is drawn around the entire line:


Goodman (1975) suggests that the beginnings of literacy occur when children
develop a knowledge that written languages makes sense and desire to understand
how it makes sense. As they begin to explore the many forms of print, they develop
concepts about print which Goodman has termed "roots of literacy." These roots of

   You can't edit the line, but you can copy or clear it.

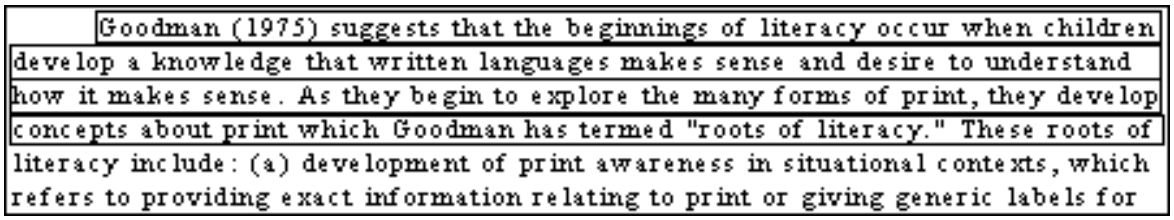
To select more than one text line at a time:

1. Select a single text line as described in the previous example.

2. Move the pointer over another text line.

3. Hold down the **Ctrl** key and the Shift key and click anywhere in the text line (**Ctrl-Shift-click**).

Each time you select another line, a box is drawn around it. The previously selected lines stay selected. The lines don't have to be next to one another to be selected.


Goodman (1975) suggests that the beginnings of literacy occur when children develop a knowledge that written languages makes sense and desire to understand how it makes sense. As they begin to explore the many forms of print, they develop concepts about print which Goodman has termed "roots of literacy." These roots of literacy include: (a) development of print awareness in situational contexts, which refers to providing exact information relating to print or giving generic labels for

4. Repeat steps 2 and 3 for each additional text line you want to select.

   OR

1. Move the pointer outside all text lines.

   When the pointer is outside all text lines, it is the standard arrow pointer.

2. Click the mouse button and drag diagonally.

   When you click, all previously selected text lines are deselected.

   As you drag, a dotted outline is drawn. When a text line falls within the dotted outline, it is highlighted.

3. Release the mouse button.

   All text lines that were highlighted are selected.

To deselect one or more text lines while keeping the rest selected:

1. Move the pointer over a selected text line and **Ctrl-Shift-click**.

   The text line you clicked in is deselected, but all other selected text lines stay selected.

2. Repeat step 1 for each additional text line you want to deselect.

To select all text lines:

- Choose **Select All** from the **Edit** menu.

If you are not currently editing a line, all text lines on the current page are selected. Otherwise, if the I-beam pointer is in a text line, all text in the line you are editing is

selected.

To deselect all text lines:

1. Move the pointer outside all text lines.

   When the pointer is outside all text lines, it is the standard arrow pointer.

2. Click the mouse button.

   All selected text lines are deselected.

To delete one or more text lines:

1. Select the text lines to be deleted.

2. Press **Delete** to remove the selected text lines.

   OR

   Choose **Clear** from the **Edit** menu to remove the selected text lines.

   A message appears asking you if you want to delete the text.

3. Click **OK**.

   All selected lines are removed from the page. The remaining lines do not close up.

To copy one or more text lines:

1. Select the text lines to be copied.

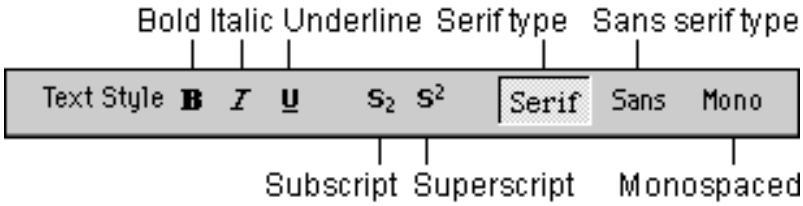
2. Choose **Copy** from the **Edit** menu to copy the selected text lines.

   All text from the selected lines is copied to the clipboard.

To apply text styles to one or more text lines:

1. Select the text lines to change.

2. Choose the style to be applied to the selected text lines from the **Style** menu or click the button on the **Style** ribbon.



You can only apply a text style in the text view. You may apply a text style to any selected text. Text can be styled with any combination of Bold, Italic, and/or Underline. All text from the selected lines is changed to the selected style.

3. Repeat Step 2 for each additional style to be applied.

4. After pages have been edited, you can save your changes in a Pro OCR file format so you can edit the pages later on from within Pro OCR, or you can save directly to a supported output file format. For more information, see Chapter 6, "Saving and Printing Documents."

Using Dictionaries in Pro OCR

To improve recognition accuracy, Pro OCR performs automatic, internal spelling verification during the Recognize step. This automatic spelling verification helps Pro OCR identify suspect characters in the scanned text. Pro OCR does this using its General dictionary, and—if you choose one—a user dictionary. There are three types of dictionaries:

- The General dictionary is an English-language dictionary. It comes with Pro OCR and is used automatically during the recognition process.

- A user dictionary is a file that contains words you've added to it that aren't in the General dictionary. Usually, you'll create a user dictionary with proper names, technical terms, product terminology, and other specialized words not included in an ordinary dictionary, so that Pro OCR will be able to use these words to help identify characters during the recognition process.

When you install Pro OCR, the General dictionary and the default user dictionary, `USER.DIC`, are automatically installed.

Checking Spelling in a Document

When you use Proof with the Misspelled Words proofing option selected, Pro OCR searches for words that are not in the General dictionary, the current user dictionary, or any supplemental Pro OCR dictionary in the dictionaries directory. Pro OCR selects the first candidate word it finds after the insertion point or the start of the current page. You can inspect the image of the selected word with the On-Screen Verifier, edit the word if necessary, add it to a user dictionary, or use Find & Replace to find additional instances of the same word.

You can check the spelling in a document at any time after it has been recognized. You can check spelling only in the text view.

NOTE: Selecting a word does not reduce the number of suspect or illegible characters in the word. You correct suspect or illegible characters by typing over them.

Sometimes, a word that Pro OCR identifies as a misspelled word is in fact a real word that just is not in the dictionary (for example, your product's trade name). By asking Pro OCR to find these words, you can easily add the word to the current user dictionary by choosing Add to User Dictionary from the Edit menu while the word remains selected.

After you've added the word to your user dictionary, the next time Pro OCR recognizes a page and finds this word, it can be used to help identify any suspect characters it may contain. In this way, you can help Pro OCR to be even faster and more accurate on documents that contain specialized words, by adding them to your user dictionary.

TIP: You can also right click on the selected word to add it to the user dictionary.

Adding Words to a User Dictionary

The simplest way to maintain a user dictionary in Pro OCR is to add words to the default user dictionary, `USER.DIC`. Unless you change it, `USER.DIC` will always be open when you open Pro OCR.

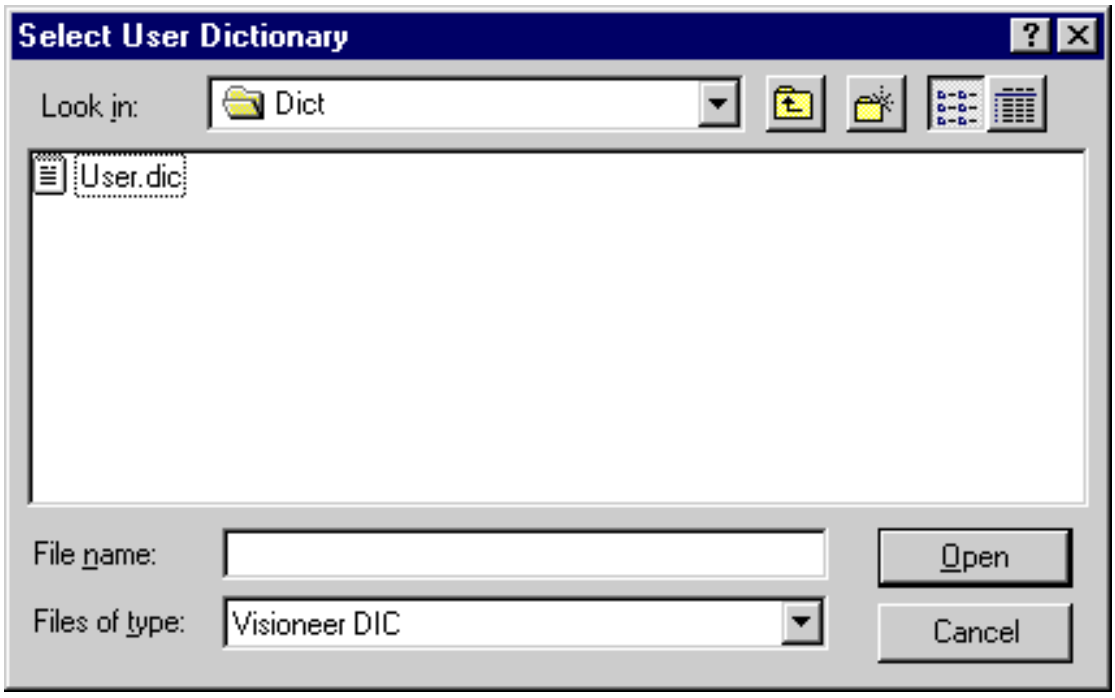
If you want to have more than one user dictionary, you can create and name more user dictionaries and then select which dictionary you want to use. The current user dictionary is automatically saved when you exit Pro OCR and when you open a different user dictionary.

You can locate a user dictionary anywhere on your hard disk and its location is remembered by Pro OCR. If you move it, you'll have to tell Pro OCR where to find it the next time you open Pro OCR.

To create a user dictionary:

1. Choose **Select User Dictionary** from the **Tools** menu.

   The following dialog box appears:



2. Type in the name of the new dictionary.

3. Click **OK**.

   The new dictionary is created and automatically selected.

To select a user dictionary:

1. Choose **Select User Dictionary** from the **Tools** menu.

   The Select User Dictionary dialog box appears.

2. Find the dictionary you want to open and select it.

   By default, the user dictionary is stored in the DICT folder. Only the dictionaries that Pro OCR recognizes appear.

   NOTE: A Pro OCR user dictionary is a simple text file of words separated by carriage returns. Any text file of this form should be usable in Pro OCR as a user dictionary. For example, you can use a word processor to create the new dictionary and then choose it for use with Pro OCR.

3. Click **OK**.

   The current user dictionary (and any changes you make to it) is used until you choose a different one.

To add to the User Dictionary while editing in the text view:

1. Select a different user dictionary, if necessary, by choosing **Select User Dictionary** from the **Tools** menu.

   NOTE: Make sure you have a user dictionary open. Add to User Dictionary is only available when there's a current user dictionary.

2. If you're searching for possible misspelled words, choose **Proof** to skip to the next possible misspelled word.

   Pro OCR selects the word. (Make sure you have the Misspelled Words proofing option selected.)

   OR

   Double-click a word to select it.

3. Choose **Add to User Dictionary** from the **Tools** menu.

   Pro OCR adds the word to the current user dictionary.

   The changes you make to the current user dictionary are automatically saved when you choose a different user dictionary or when you exit Pro OCR.

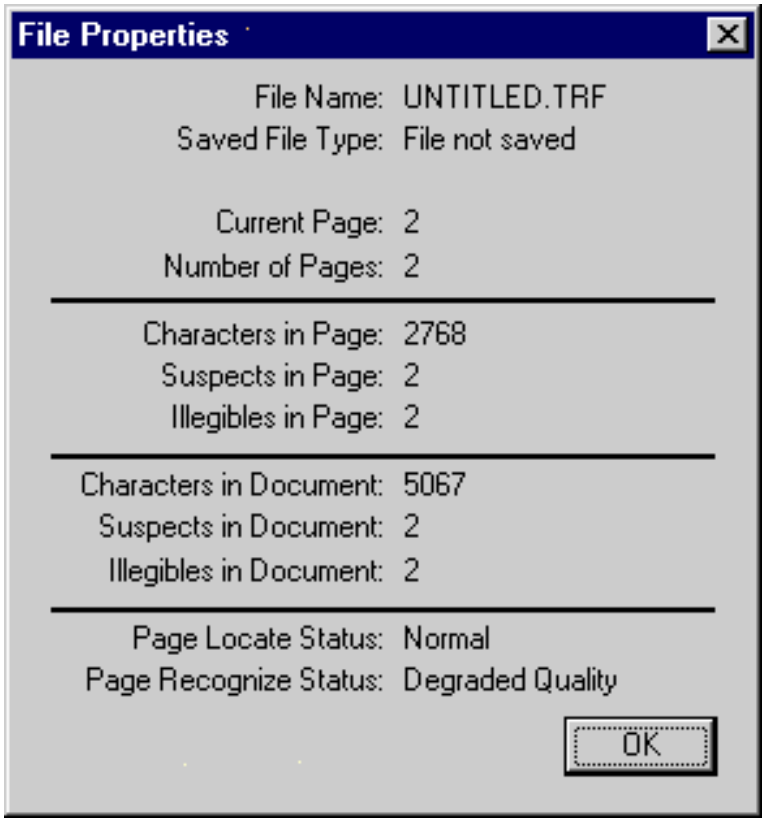
Displaying a Summary of Recognized Errors

When recognition is completed, you can display the Properties dialog box to view file information, such as the number of pages, number of characters, and number of suspect characters.

To display summary information:

- Choose **Properties** from the **File** menu.

  The File Properties dialog box appears.



© Copyright 1998 Visioneer, Inc. Reach us at www.visioneer.com.

Contents

Chapter 3: Getting Documents

Getting a Page—The Basic Steps

Getting Pages From a Scanner

Setting Scanning Options

Selecting a Scanner as the Source

Getting a Page Using a Scanner

Using Auto OCR with Scanners

Getting Pages from an Image File

Selecting a File as the Source and Getting Pages

Getting Files From Other Scanner Applications

Getting Fax-modem Files

Using Auto OCR With a File

More About Enabling Auto OCR Dialogs

Glossary

Contents

Chapter 4: Locating Text and Graphics

Kinds of Locate Regions

Text Regions

Numeric Regions

Picture Regions

Tables

Pro OCR's Locating Methods

Locating Text and Pictures

Locating with a Template

Order of Locate Regions

Examples of Locating Documents

Processing Resumes

Processing Legal Documents

Processing Faxed Documents

About Columns, Locate Regions, and Output File Formats

Defining Locate Regions Manually

Tips When Creating Locate Regions

Overlapping Locate Regions and Skewed Text

Selecting and Deselecting Locate Regions

Changing the Kind of a Locate Region

Deleting a Locate Region

Resizing a Locate Region

Reordering Locate Regions

Glossary

Contents

Chapter 5: Setting Recognize Options and Proofing a Recognized Document

Selecting Type Quality Options

Selecting Display Options

Setting the Suspect Character Threshold

Setting the Illegibles Character Symbol

Selecting a Display Font

Indicating Whether Pictures Appear During Text View

Recognizing a Single Page

Working with Recognized Pages in Text view

Setting the Zoom Levels

Selecting a Page to Display

Selecting Text or Image View

Proofing

Selecting Proofing Options

Proofing a Document

Reviewing and Editing Text in the Text View

Using Dictionaries in Pro OCR

Checking Spelling in a Document

Adding Words to a User Dictionary

Displaying a Summary of Recognized Errors

Glossary

Contents

Chapter 6: Saving and Printing Documents

Saving Documents and Other Pro OCR Items

Saving a Document

Saving Templates

Saving Settings

Supported Output File Formats

Saving to Proprietary Pro OCR Formats

Saving to Standard Image File Formats

Saving to Generic Text File Formats

Saving to Application Formats

Format Suppression and Customizing

Exporting to a Word Processor that Pro OCR Doesn't Support

Saving Pictures

Printing a Document

Glossary

Contents

Chapter 7: Creating and Processing Deferred and Batch Jobs

The Advantages of Finish and Deferred Processing

Guidelines for Using Finish Processing and Deferred Processing

How it Works

Setting Up and Processing Deferred Jobs

Processing Deferred Jobs

Batch Processing

Glossary

Chapter 7

Creating and Processing Deferred and Batch Jobs

This chapter tells you how to process Deferred, Finish, and Batch jobs.

Finish Processing lets you combine the efficiency of multi-step automatic operation with the power and flexibility of single-step interactive operation. You can process pages in your document according to their specific characteristics, while still having automatic processing available for the rest of the pages in the document.

The two stages of deferred processing—Create Deferred Job and Process Deferred Jobs—give you similar efficiency and flexibility, and additionally let you save the documents and automatically complete processing from the saved file.

With Batch Processing you can specify a source directory that contains image files, and then process the files that are the same type all at the same time. The processed files are automatically saved in a format that you select to a specified destination directory.

The Advantages of Finish and Deferred Processing

When you use automatic processing (Auto OCR), you can efficiently and automatically perform Get Page, Locate, and Recognize on large stacks of pages. However, you can only use one set of Gallery settings on that document during an automatic processing session. Automatic processing is fine when the same settings are appropriate for all pages. But when you need to apply different settings to different pages and still want the ability to process large numbers of pages at once, use finish processing or deferred processing.

Finish Processing and the two stages of deferred processing let you fill in the gaps to handle the pages that one set of Gallery settings won't handle. Perform this type of processing when some of the pages you're processing need more individual attention than Auto OCR can provide, but you don't want to be tied to single-step processing of every page. For example, when you need more than one Get Page,

Locate, or Recognize setting for the different pages in a document. You'll also use these processes with a mixture of settings or when more than one person works on the documents or more than one workstation is used.

Guidelines for Using Finish Processing and Deferred Processing

You can combine automatic processing, single-step processing, Finish Processing, and deferred processing in a variety of ways:

- Process a whole document using automatic processing.
- Process a whole document using automatic processing, then use single-step procedures to process individual pages again as necessary.
- Read in a whole document using Create Deferred Job, then use single-step procedures to process individual pages as necessary, and complete processing automatically with Finish Processing.
- Read in a whole document using Create Deferred Job, then use single-step procedures to process individual pages as necessary, and save the document in Pro OCR Deferred format. You can then complete processing automatically with Process Deferred Jobs.

How it Works

When you select Finish Processing, you're telling Pro OCR to intelligently evaluate the current document and automatically complete processing. When you select Process Deferred Jobs, you're telling Pro OCR to read in and intelligently evaluate a saved document and automatically complete processing.

As Pro OCR encounters each page in the document, it checks to see if the page has been located or recognized. If the page has not yet been located or recognized, Pro OCR uses the current Gallery settings for the Locate and Recognize steps. If the page has been located but not recognized, Pro OCR uses the already-specified locate regions for the page and uses the current recognize setting for the Recognize step. If the page has already been located and recognized, Finish Processing or Process Deferred Jobs continues with the next page.

Setting Up and Processing Deferred Jobs

Use Create Deferred Job to get pages and save them in the Pro OCR Deferred format for processing later on. After you create a deferred job, you can use any combination of locating and recognizing on some or all pages and then save the document. When you're ready to finish processing the saved document, use Process Deferred Jobs to automatically perform any additional processing.

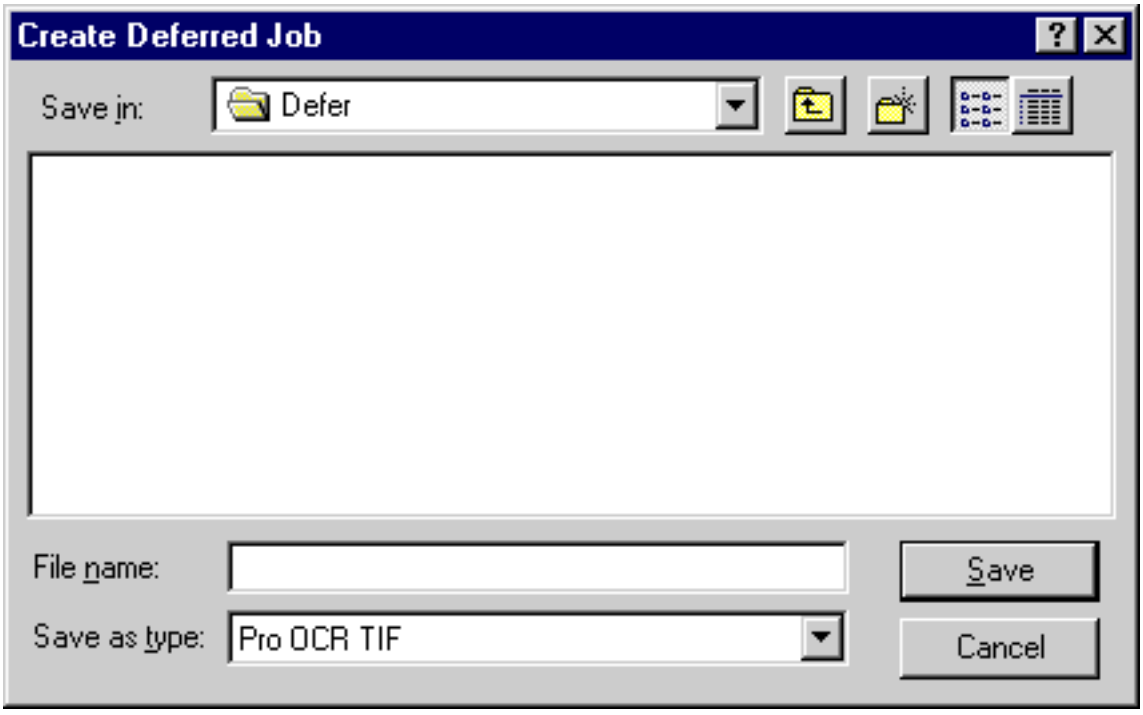
To create a deferred job:

1. Select **Use Scanner** or **Open File** from the **Get Page** drop-down list in the Gallery toolbar.

   You can create a deferred job either by scanning pages or by reading them in from a file. If your source is a scanner, don't forget to specify the appropriate scanner settings.

2. Choose **Create Deferred Job** from the **Recognize** menu.

   The Create Deferred Job dialog box appears.



3. Change directories, if necessary.

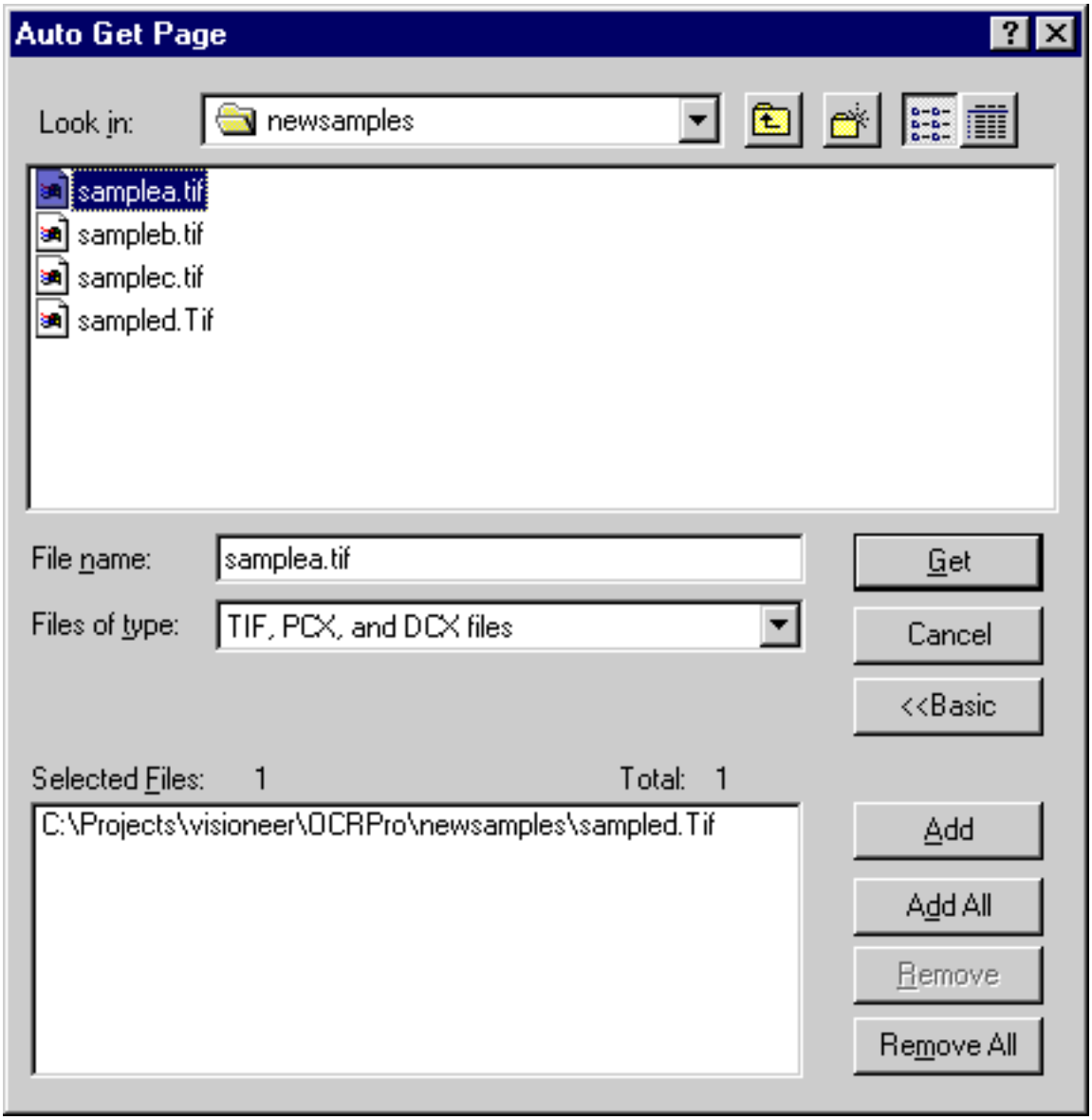
   When you choose Process Deferred Jobs, Pro OCR opens the DEFER directory by default. If you save to the DEFER directory, you won't have to search through the directory hierarchy to find the file later on. However, Process Deferred Jobs can open files saved in the Pro OCR Deferred format

from any directory or disk.

4. Type in a new file name.

5. Click **Save**.

   If Open File is selected in Get Page as the source to get pages from, the Auto Get Page dialog box appears. If Use Scanner is selected as the source, Pro OCR immediately starts to scan.



6. Scan the documents, or if your are getting a file, select a file in the Auto Get Page dialog box, and then click the **Get** button.

   To select multiple files, click the **Advanced** button, choose a file, and click **Add**. Repeat this process until you select all files that you want to get, then

click the **Get** button.

The Get Page process is the same as when you're using Auto OCR with either a scanner or a file.

7. When you're finished getting pages, click **Finished**.

   The pages are read in the same way that they are when you use Auto OCR. When all pages are read in, a dialog box tells you the process is completed.

8. Click **OK**.

   The pages are saved to the file you named previously. The last page of the document is displayed at the last selected zoom level in the image view.

9. You can continue by processing individual pages, or you can complete processing now by choosing **Finish Processing** from the **Recognize** menu or later with by choosing **Process Deferred Jobs** from the **Recognize** menu.

Processing Deferred Jobs

Use Process Deferred Jobs to complete the processing of files saved in the Pro OCR Deferred format.

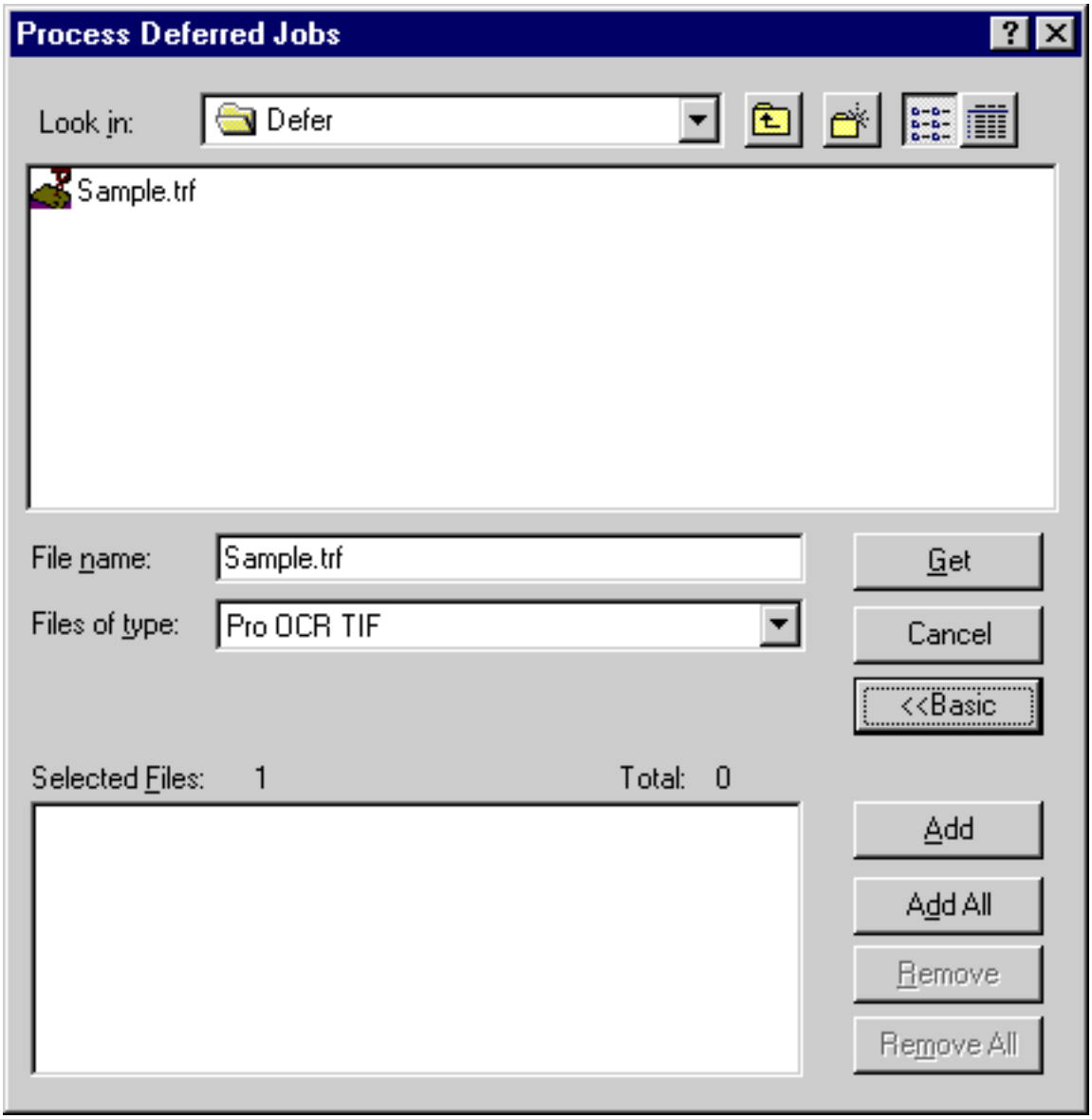
To process a deferred job:

1. Select options in the **Locate** and **Recognize** drop-down lists in the Gallery toolbar.

   Any pages that don't already have locate regions defined or have not been recognized are processed based on the current Gallery Toolbar selections.

2. Choose **Process Deferred Jobs** from the **Recognize** menu.

   The Process Deferred Jobs dialog box appears.



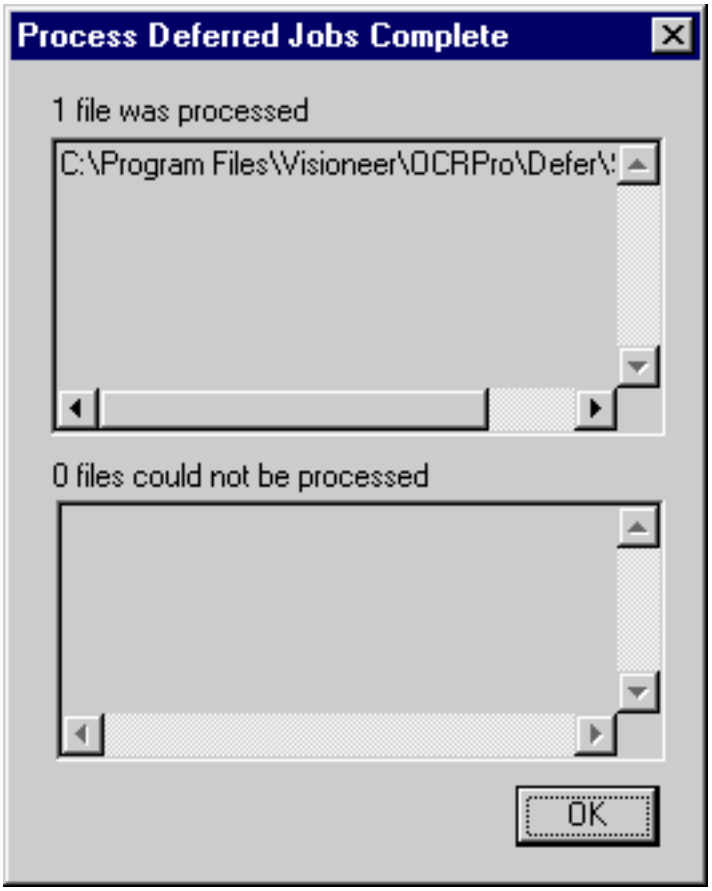
Deferred jobs are saved in the Pro OCR Deferred format with image, locate regions (if any), and recognized text (if any).

3. Select the file you want to process and click **Get**.

   To select multiple files, click the **Advanced** button, choose a file, and click **Add**. Repeat this process until you select all files that you want to get, then click the **Get** button.

   Pro OCR reads the deferred and locates or recognizes any regions that were not previously located or recognized.

   When all processing is done, the following dialog box appears.



NOTE: The Process Deferred Jobs command does not process non-Pro OCR image files. If some of your files could not be processed, read them in again (using Get Page, Auto OCR or Create Deferred Job) and process them as you normally would.

The last page of the document is displayed at the last selected zoom level in the text view.

4. Click **OK**.

   You can now proof the document (press Tab), and edit it as needed.

Batch Processing

Use Batch Process to convert all of the files that are of the same image type (such as TIFF) in a specific directory at the same time. For example, if you have a stack of invoices that you scanned and saved as TIFF files, you can process all of the invoices at the same time.

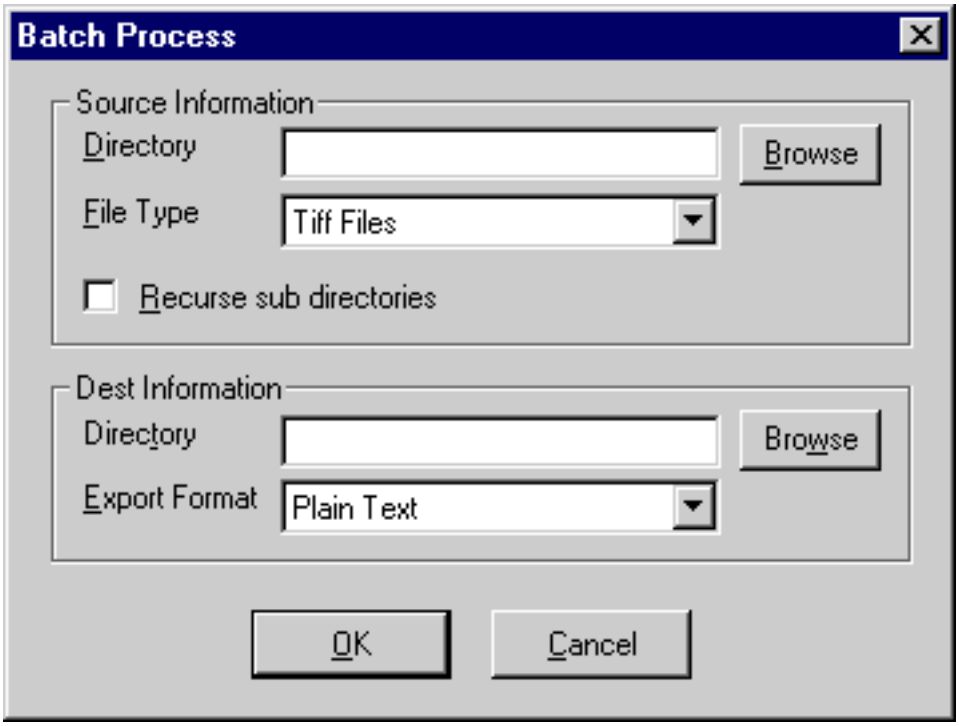
Batch Process allows you to specify the source directory that contains image files, image file type, destination directory where the recognized results are saved, and the export Format. Pro OCR automatically performs the OCR job on each image file under the source directory, and exports the results to the destination directory.

To process as a batch:

1. Select **Locate** and **Recognize** options from the drop-down lists in the Gallery toolbar.

2. Select **Batch Process** from the **Recognize** menu.

   The Batch Process dialog box appears.



3. Choose a file type from the Source Information **File Type** drop-down list.

   This is the type of files you want to process, such as TIFF.

4. Click the Source Information **Browse** button and choose the source directory.

   All the files in this directory that are of the image type you selected in the previous step will be processed.

5. Choose an export format from the Destination Information **Export Format** drop-down list.

The export format determines the saved format and the extension of all of the files included in this Batch Process. Batch Process names each file by combining the file name of the image and the default extension name of the export format. For example, if the image file's name is sample.tif, and you choose Plain Text as the export format, the result file is sample.txt.

The following is a list of default extension name of all the export formats supported in Batch Process.

- TXT—for Plain Text, Text with Line Breaks, Comma Delimited Text, Formatted Text, and Tab Delimited ASCII
- SAM—Lotus Ami Pro
- WK1—Lotus 1-2-3
- XLS—Microsoft EXCEL
- DOC—Microsoft Word
- RTF—Rich Text Format
- WPF—WordPerfect
- HTM—Hyper Text Markup Language

6. Click the Destination Information **Browse** button and choose the destination for the processed information.

   All processed files are saved to this location.

7. Click **OK** to start the Batch Process.

   The progress of Batch Process is shown on the Title Bar of Pro OCR window. Each processed file appears in the destination directory that you specified.

© Copyright 1998 Visioneer, Inc. Reach us at www.visioneer.com.

Contents

Chapter 8: Tips for Getting the Best Results

Fixing Broken and Touching Characters

Adjusting Brightness for Consistent Documents

Handling Documents That Are Not Consistent

Processing Documents with Different Page Sizes or Orientations

Processing Documents with Different Character Quality

Converting Parts of a Page in a Multipage Document

Changing the Gallery Options

Using Get Page Again

Using Locate Again

Using Recognize Again

Finding and Replacing Recognized Text

Making Sure Page Images are not Skewed

Using Numeric Regions When You’re Recognizing Numeric Text

Putting Pages in the Scanner Properly

Avoiding Markings on Pages

Glossary

Chapter 8

Tips for Getting the Best Results

This chapter provides tips for getting the best results from Pro OCR by:

- Fixing broken and touching characters
- Adjusting the brightness to obtain consistent documents
- Processing inconsistent documents
- Changing a setting after completing autoprocessing
- Getting the best recognition
- Making sure page images are not skewed
- Using numeric regions when you're recognizing numeric text
- Putting pages in the scanner correctly
- Avoiding marks on a page

Fixing Broken and Touching Characters

Pro OCR is good at recognizing characters that are broken (light) or touching (dark), especially when you use brightness level to compensate for poor character quality.

You can assist Pro OCR in accurately recognizing text that has broken/light or touching/dark characters by adjusting the brightness level used during scanning. There are two general rules:

- When characters are dark or touching, use a higher (brighter) setting.

- When characters are light or broken, use a lower (darker) setting.

However, when there are both broken and touching characters on the same page or in the same document, trying to fix one problem may make the other problem worse.

In such a situation, you'll usually find that it works to use this rule:

- When there are both broken and touching characters, use a lower (darker) setting—that is, fix the broken characters.

It's usually better to decrease the brightness level (darken the image) to compensate for the broken or light characters, even though by doing so you may increase the number of dark or touching characters.

Because light or broken characters are more of a problem than dark or touching characters, decrease the brightness just enough to compensate for the broken characters.

Adjusting Brightness for Consistent Documents

For most documents, you'll find that using Auto OCR works well. Auto OCR is most useful when the pages in your document are consistent:

- The same page size and orientation
- The same printing source
- The same photocopy generation (that is, how many times the page has been recopied). The quality of the character image degrades each time you make a photocopy of a photocopy (that is, a "second generation photocopy").

When you adjust the brightness setting in your scanner software to compensate for poor photocopies, you may find that different "generations" of photocopies need different brightness settings.

To find the correct Brightness setting:

1. In your scanner software, increase or decrease the brightness setting.

You may have to experiment with different settings.

If your scanner supports Auto brightness, you may want to try it first before setting brightness manually.

2. Click the Get page button in the Gallery, or choose **Get Page** from the **Process** menu.

3. Get the file that you want to adjust.

4. Zoom in on the page to check the image's quality.

   A good image has characters that are not too dark or touching and are not too light or broken. If the image looks good, skip to step 10.

5. If you're not sure if it's a good image, use Locate and Recognize on the page and check the results.

   If the page has many suspect and illegible characters in the recognized text, you may be able to improve recognition by changing the Brightness setting.

6. If you want to scan the page again with a different brightness setting, delete the page by choosing **Delete Page** from the **Edit** menu.

7. Increase or decrease the brightness setting in your scanner software.

   If the page image contains dark and/or touching characters, increase the brightness. If the page image contains light and/or broken characters, decrease the brightness.

8. Use **Get Page** again on the same page.

9. Repeat Steps 3 through 8 until you get the image that you want.

10. When you have an appropriate setting, delete any extra pages, then process all the pages in the document using Auto OCR.

NOTE: When all of the pages in a document have a consistently "noisy" (fuzzy, dotty) background (as in some multi-generation photocopies and some faxes), or are on the same colored background or paper, you'll increase brightness to "fade out" the background "noise." You'll have to be careful not to increase it so much

that you begin to make the characters too light and/or broken.

Handling Documents That Are Not Consistent

Sometimes the pages in your document are not consistent, for example, do not have the same page size. To handle this, you must change the Gallery options for each page. In such cases, use Get Page, Locate, or Recognize in combination with Process Deferred Jobs or Finish Processing. The steps or combinations of steps you use depend on the characteristics of each document.

To fix documents that aren't consistent:

1. Determine which settings in the Gallery toolbar apply to the most pages.

2. Use the **Get Page, Locate,** or **Recognize** commands on the pages that need to be processed with different settings.

3. Set the controls in the Locate and Recognize rows of the Gallery, as determined in step 1, for the rest of the pages.

4. Choose **Finish Processing** from the **Process** menu.

   OR

   Save the document in the Pro OCR Deferred format.

   When you use either Finish Processing or Process Deferred Jobs, Pro OCR checks each page to see if it has already been located and recognized. If a page has been located, Pro OCR uses the existing locate regions. If a page has not been located, Pro OCR applies the current locate settings. If a page has been recognized, Pro OCR skips it. If the page hasn't been recognized, Pro OCR applies the current recognize settings.

Processing Documents with Different Page Sizes or Orientations

You can process a document that has mixed page sizes, for example, US Letter or US Legal, or different orientations (portrait and landscape).

To process a document with mixed page sizes or orientation:

1. Select the appropriate page orientation in your scanner software for the page

to be processed.

2. Click **Get Page** to get the page, or choose **Get Page** from the **Process** menu.

3. Repeat Steps 1 and 2 for each page in the document.

4. Choose **Finish Processing** from the **Process** menu.

   Make sure you select the Locate and Recognize options in the Gallery toolbar.

   OR

   Save the document in Pro OCR Deferred format.

   If you save the document in the Pro OCR Deferred format, you can finish processing it by choosing **Process Deferred Jobs** later.

Processing Documents with Different Character Quality

A good image has characters that are not too dark or touching and are not too light or broken. It also has characters that are distinct from the background. The background in a good image is light and not “fuzzy” or “dotty.”

To process a document with pages that vary in character image quality (too dark/touching, too light/broken) or in background (color or “noise”):

1. Use the brightness level control in your scanner software to manually select a brightness level.

2. Click **Get Page** to get the page, or choose **Get Page** from the **Process** menu.

3. Zoom in on the page to see if it’s a good image.

   If the image looks good, skip to step 9.

4. If you’re not sure if it’s a good image, use the Locate and Recognize on the page, then check the results.

   If there are many suspect and illegible characters in the recognized text, you may be able to improve recognition by experimenting with a different brightness setting.

5. To scan the page again with a different brightness setting, delete the page.

6. Increase or decrease the brightness.

   If the page image contained dark or touching characters, increase brightness. If the page image contained light or broken characters, decrease brightness.

   If the page image has a “noisy” (fuzzy, dotty) background (as in some multi-generation photocopies and some faxes), increase brightness in order to “fade out” the background “noise.”

7. Repeat steps 2 through 6 until you get the image that you want for that page.

   After you find the correct setting for one page, you can use this setting for other pages in your document that are of similar quality.

8. Choose **Finish Processing** from the **Process** menu.

   Make sure you set the appropriate Locate and Recognize options.

   OR

   Save the document in the **Pro OCR Deferred** format.

   If you save in the Pro OCR Deferred format, you can choose Process Deferred Jobs at a later time.

NOTE: When you use Get Page after deleting a page, the new page is inserted after the current page. If you delete a page and it is not the last page of the document, make sure you go to the page preceding the deleted page before you get the page again.

Converting Parts of a Page in a Multipage Document

This procedure shows how to process each page separately at the Locate step, but you can use the same Get Page and Recognize options for the entire document.

Creating the Deferred Job

To process a document when you want different information from each page:

1. Choose **Create Deferred Job** from the **Process** menu.

   The Create Deferred Job dialog box appears.

2. Select the files you want to process.

   Create Deferred Job lets you scan a stack of pages or read in a set of image files. If you have only one page, or you want to retrieve and process a single file, you can use Get Page.

   Remember to set the appropriate Get Page options.

3. Manually locate the locate regions on the current page.

4. Repeat Step 2 for each page in the document.

5. Choose **Finish Processing** from the **Process** menu.

   Make sure you set the appropriate Locate and Recognize options in the Gallery.

   OR

   Save the document in the Pro OCR Deferred format.

   If you save the document in the Pro OCR Deferred format, you can choose Process Deferred Jobs later to finish processing it.

Using Locate and Recognize on the Document

This section discusses documents where some pages need a different locating method, but all the remaining pages can be located with the same locating method. For example, if you need to locate text and pictures on some pages or text only on other pages, you can change the locating method. This procedure assumes you'll use the same settings in the Get Page and Recognize rows of the Gallery for the entire document.

To use Locate and Recognize on a document and then complete processing:

1. Choose **Create Deferred Job** from the **Process** menu.

   Remember to set the appropriate Get Page options.

2. Determine which Locate and Recognize options in the Gallery apply to the majority of pages.

   For example, the locating options might be Locate Text Only and Single Columns Only. You'll use these settings in step 4.

3. Use Locate and Recognize, as necessary, on each of the other pages.

   In other words, you'll leave the "majority pages" alone and only use Locate—or Locate and Recognize—on the pages that are not "majority pages."

4. Choose **Finish Processing** from the **Process** menu.

   Make sure you set the Locate and Recognize settings that you decided on in Step 2. You'll use these settings to process all the "majority pages" (that is, the pages you didn't use Locate and Recognize on in step 3).

   OR

   Save the document in the Pro OCR Deferred format.

   If you save the document in the Pro OCR Deferred format, you can use Process Deferred Jobs at a later time to finish processing it. When you do, make sure you set the controls in the Locate and Recognize rows of the Gallery the way you decided on in step 2.

Each of the above procedures shows you how to handle one particular situation. Note that you may combine these procedures when appropriate to handle combinations of situations.

Changing the Gallery Options

Sometimes, after using Auto OCR on a document, you'll find that you've chosen an inappropriate Gallery option for one or more steps or pages. When this happens, you don't have to start all over again. With Pro OCR, you can redo only the steps or pages that you need to. The rest of your document is not affected. The following scenarios give you some hints and suggestions about using Get Page, Locate, and

Recognize over again.

Using Get Page Again

You may want to use Get Page again if you scan pages in with an incorrect Page Size or Orientation setting, or if you didn't use an appropriate brightness or scanning resolution setting. You can use Get Page again at any step in the Pro OCR process.

If you get a page, the page is added after the current page. If the page's quality is not good, you can delete it and redo the steps.

NOTE: When you use Get Page after deleting a page, the new page is inserted after the current page. If the page you deleted was not the last page of the document, make sure you move back to the page preceding the deleted page each time you repeat this step.

Using Locate Again

This may be necessary if you decide that a located page has incorrect locate regions on it, or if you change your mind about whether or not to locate picture regions. You can use Locate again at any step in the Pro OCR process.

To locate the current page again:

1. Select the appropriate Locate options from the Locate drop-down list.

2. Click the Locate button, or choose **Locate** from the **Process** menu.

   OR

1. Delete some or all locate regions, if necessary.

2. Manually locate new locate regions, or resize existing locate regions, or redefine existing locate regions, if necessary.

   You may use Locate again for individual pages in the current document.

NOTE: After you locate a page again, you must use Recognize on a page again before any changes to the locate regions show up in the recognized text.

Using Recognize Again

This may be necessary if the text on a page was not recognized accurately because of an incorrect type quality setting. You can recognize again at any step in the Pro OCR process.

To recognize the current page again:

1. Select the appropriate Recognize options from the Recognize drop-down list.

2. Click the Recognize button, or choose **Recognize** from the **Process** menu.

   You can use recognize again for individual pages in the current document.

Finding and Replacing Recognized Text

If Pro OCR incorrectly identifies a character in one place in the document, it may incorrectly identify the same character everywhere. For example, the document may be a monthly sales report from the XYZ Company, typed on a typewriter with a broken “X” that Pro OCR couldn’t identify. While you’re proofing the document using Proof, with the Illegible Characters proofing option selected, you notice that Pro OCR has substituted the currently chosen Illegible Character Symbol—“@”—everywhere that it encountered the broken “X.” You can easily change all the occurrences of “@YZ” to “XYZ” using Find & Replace.

You use Find & Replace to search for and replace repeated occurrences of text. The text you search for can be specified in several ways:

- It can be text that you type in the Find & Replace dialog box.

- It can be text that you’ve selected manually in the document.

- It can be text that Pro OCR has selected during Proof.

To use Find & Replace with Proof:

1. Choose **Options** from the **Tools** menu.

   The Options dialog box appears.

2. Click the Proofing tab, and select the following options: **Suspects (Normal)**, **Illegibles**, and **Misspelled Words**.

3. Click **OK**.

4. Choose **Proof** from the **Process** menu.

   When Proof selects the character to replace, select the word that contains the suspect character or illegible character you want to replace.

5. Choose Find & Replace from the Edit menu.

   The dialog box is displayed with the selected text.

6. Type the correct text in the Replace box.

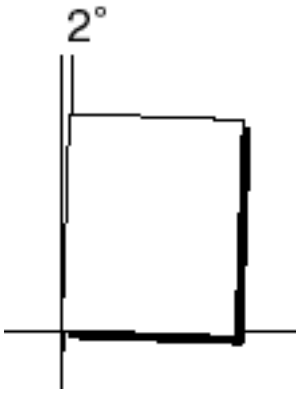
7. Click the “Replace then Find” button.

   The current occurrence is replaced.

8. Continue clicking “Replace, then Find” until you’ve changed all occurrences of the current Find text.

NOTE: If you want to change the same text throughout the document, you can click Replace All once instead of clicking “Replace, then Find” over and over again. The Replace All operation cannot be undone.

Making Sure Page Images are not Skewed



Even with good quality characters on good quality paper, Pro OCR will have trouble locating and recognizing accurately if the type in the page image is skewed (crooked). This can happen either because the text is crooked on the page or because the page is scanned at an angle.

What's important is that the text image cannot be skewed more than 2° for Pro OCR to have accurate recognition. The illustration to the left shows a page that has 2° skew.

If text is skewed (at an angle) on the page, both locating and recognizing may be affected. When text on a page is badly skewed, Pro OCR may have trouble correctly locating paragraph boundaries and recognizing the contents of these paragraphs. There are two ways to fix this problem:

- You can adjust the paper so that the text is scanned in straight, or is not skewed more than 2° . If text is straight on the page, make sure that the paper is put in the scanner straight.

- You can select the Straighten Skewed Images processing option. When you select Straighten Skewed Images, Pro OCR can automatically rotate the image of the page up to 15 degrees in order to straighten the text on the page.

Using Numeric Regions When You're Recognizing Numeric Text

Use numeric regions for text in your document that will only consist of numbers. Pro OCR will make sure that all characters in the numeric region are recognized as numbers and not mistaken for letters.

Putting Pages in the Scanner Properly

Make sure that you put pages in the scanner with the proper orientation—Portrait or Landscape—and use the corresponding Get Page setting.

Avoiding Markings on Pages

Handwritten notes on pages may slow down recognition. You can reduce the effect of markings on pages by:

- Scanning the document first and then marking it up, or making a photocopy for scanning before you mark it up.

- Using whiteout to remove any markings that don't overlap text. Be very careful, however, about using whiteout on text—you may make the text even more illegible. If you don't want to mark up your original document, make a photocopy and use whiteout on it.

© Copyright 1998 Visioneer, Inc. Reach us at www.visioneer.com.

Index

A

accuracy of recognition
ADF and Auto OCR
All Pages in One File (Split Document options)
application formats for saving
Auto Get Page dialog box (1)
Auto Get Page dialog box (2)
Auto OCR
    from a file
    from a scanner with a flatbed
    with an ADF scanner
auto orientation

B

Batch Process dialog box
batch processing
    explanation
    selecting
brightness, adjusting
broken characters, fixing

C

character quality, processing different
Create Deferred Job dialog box

D

DCX file format

deferred processing
advantages
continuing job
creating job
explanation
guidelines
setting up
Degraded or Fax Quality command
deleting locate regions
dictionary
adding words
creating
General
user
See also
user dictionary
directories
deferred jobs
dictionaries
discarding format when saving
Display Options command (1)
Display Options command (2)
Display Options command (3)
Display Pictures option

E

editing
all lines
applying styles
copying
deleting text
deselecting
in text view
more than one line
single line (1)
single line (2)
text
Enable Auto OCR Dialogs option (1)
Enable Auto OCR Dialogs option (2)
Enable Auto OCR Dialogs option (3)
errors, summary displayed in Get Info

exporting to unsupported word processor

F

fax example
faxed document processing
fax-modem files
features
file formats
    input
        DCX
        fax-modem files
        files from other scanner applications
        PCX
        TIFF
    output
        Pro OCR
        Pro OCR Text Only
        spreadsheet
        standard text
        word processor
File menu
    Process Deferred Job command (1)
    Process Deferred Job command (2)
    Save As command
File Properties dialog box
file, getting multiple
files from other scanner applications
finding and replacing recognized text
finish processing
    advantages
    guidelines
flatbed scanning
format suppression when saving

G

Gallery settings
    changing
    explanation

retrieving
saving
source controls (1)
source controls (2)
type quality controls
Get Info
get page
basic steps
files from unsupported scanners
from file
from scanner
getting fax-modem files
getting multiple files (1)
getting multiple files (2)
one scanned page
scanning additional pages
setting options (1)
setting options (2)
setting options (3)
single-step operation
using Auto OCR with files
Get Page dialog box (1)
Get Page dialog box (2)
Go to Page dialog box

H

hints
HTML (1)
HTML (2)

I

illegible characters
Image View icon
image view, selecting
ISIS upgrade

L

legal document processing
locate regions
- changing the kind of
- defining manually
- defining the order
- deleting
- kinds of
- legal document example
- locating manually
- method to use
- numeric
- order of
- overlapping regions and skewed text
- overlapping text and pictures
- picture
- redefining
- reordering
- resizing
- resume example
- selecting and deselecting
- single or multiple columns
- tables
- text
- text and pictures (1)
- text and pictures (2)
- tips
- using a template

M

magnifying the view
misspelled words

N

normal suspect threshold

numbers and alphanumeric words
Numeric Region icon
numeric regions

O

One Page Per File option (Split Document options)
On-Screen Verifier
- example of use (1)
- example of use (2)
- showing in Text View
- turning on or off

opening a file
Optical Character Recognition (OCR)
- defined
- uses for

Options dialog box
- Display options
- Process options (1)
- Process options (2)
- Proof options

order of locate regions
overlapping text and pictures

P

Page controls (Status bar) (1)
Page controls (Status bar) (2)
Page Image Rotation commands
pages
- displaying in image view
- displaying in text view
- processing different orientations
- processing different sizes
- selecting to display
- zooming

PaperPort
- using to start Pro OCR
- using with Pro OCR (1)
- using with Pro OCR (2)

PCX file format
Picture Region icon
picture regions
    defined
    white out text
pictures, saving
preserving format when saving
printing
Pro OCR file format (1)
Pro OCR file format (2)
Pro OCR Text Only file format
Pro OCR window
Process Deferred Job command (File menu) (1)
Process Deferred Job command (File menu) (2)
Process Deferred Jobs Complete dialog box
Processed Deferred dialog box
processing options (1)
processing options (2)
Proof command
proofing
    combinations of characters and words
    misspelled words
    numbers and alphanumeric words
    punctuation and symbols
    selecting options
    suspect and illegible characters
    using with Find & Replace
    whole lines option
proprietary formats for saving
pull-down menus (1)
pull-down menus (2)
puncuation and symbols

R

recognition
    accuracy of
    how to get the best
    single-step operation
    speed of
Recognition Completed dialog box
resizing locate regions

resume processing
retrieve settings
rotate
RTF

S

Save As command (File menu)
Save As dialog box
Save As Options
Save As Options dialog box
saving
- a template
- as HTML
- as plain text
- as RTF
- as speadsheet
- as text
- for database
- for spreadsheet
- for wordprocessor
- Gallery settings
- multiple documents as separate files
- pictures
- pictures (example)
- to application formats
- to generic text file format
- to MS Word (example)
- to Pro OCR (example)
- to Pro OCR deferred format
- to Pro OCR format
- to Pro OCR text only
- to proprietary formats
- to spreadsheet (example)
- to standard image file format
- to word processor
- using format options
- with pictures

scanner
- selecting (1)
- selecting (2)
- using non-TWAIN compliant

scanning
- additional pages
- one page
- second side
- selecting a scanner
- setting options
- with Auto OCR and ADF
- with Auto OCR and scanner
- with flatbed

Select Source dialog box
Select Template dialog box
Select User Dictionary dialog box
selecting a scanner
single-step operation
- get page
- locate
- Recognize
- when to use

skewed images
- adjusting for
- straightening

source
- selecting file
- selecting scanner

source controls (Gallery) (1)
source controls (Gallery) (2)
speed of recognition
spellcheck
Split Document options (Save As Options)
Split on Blank Pages option (Split Document options)
splitting A3 page
starting Pro OCR
- from Start menu
- using PaperPort
- using the Wizard

Status bar
- Page controls
- View controls
- Zoom controls

straightening skewed images
Style bar
styles
suspect and illegible characters (1)
suspect and illegible characters (2)
suspect character threshold

T

- tables
  - defined
  - scanning mixed
  - single column
- template
  - creating
  - saving
  - selecting
  - using (1)
  - using (2)
  - using (3)
- text
  - applying styles
  - copying
  - deleting
  - deselecting
  - regions
  - selecting all lines
  - selecting more than one line
  - selecting single line
  - Text Region icon
- Text Region icon
- text view
  - editing operations
  - editing text
  - editing within a line
  - selecting
- Text View icon
- TIFF
- tips for locating
- toolbar
- tutorial
  - scanning a document using a template
  - scanning a document with mixed tables
  - scanning a document with tables
  - scanning and saving with pictures
  - scanning multi-column
  - scanning one page
- Type quality controls (Gallery)

U

user dictionary
- adding words
- adding words in text view
- creating
- selecting

V

view
- changing
- displaying pages
- zooming in and out

view controls (1)
view controls (2)
Visioneer format

W

White Out Text option (1)
White Out Text option (2)
Wizard
word processor
- exporting to unsupported
- saving to

Z

Zoom controls (Status bar)
zoom in and out
Zoom In and Zoom Out icons